The evolution of genomic anatomy: Linkage, expression and rates of evolution

Williams, Elizabeth Jane Bulkeley

Award date:
2002

Awarding institution:
University of Bath

The Evolution of Genomic Anatomy:

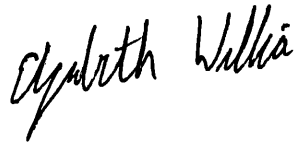
Linkage, Expression and Rates of Evolution

Submitted by Elizabeth Jane Bulkeley Williams

for the degree of Phd

of the University of Bath

2002



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

"Imagination is more important than knowledge.."

Albert Einstein

Summary

How to explain the variation both between and within species has been a problem that has fascinated evolutionary biologists for some time. However there is long standing dispute as to whether the variation seen is primarily due to selection or to random genetic drift. One way of distinguishing between these two theories is by looking at the causes of variation in rates of evolution. There is a great deal of variation in how fast genes evolve; from genes that barely change over many millions of years (e.g. Histones) and those that often show substitutions after very short periods of time (e.g. immune genes). If the variation in rates of evolution were due to variation in selective pressure around the genome, we would expect to find that linked genes show similar rates of evolution. This is what was discovered when looking at the variation in evolutionary rates around the rodent genome. Alternatively, it could be that linked genes tend to be similarly expressed. Evidence strongly suggested that genes expressed in just a few tissues tend to evolve faster and also tend to cluster causing the local similarity in rates of evolution. Other results suggested that mutation rates also vary around the genome in a slightly different fashion, tending to be chromosomally specific. I also found that the estimations of mutation rates commonly used are biased by base composition and this could have important implications for how these estimations are used. In particular, I showed that the repeatability in Ks, which describes the rate of silent substitutions in a sequence, between lineages could be partially ascribed to orthologous sequences having similar base composition. In addition, the occurrence of gene rearrangements seems to be affected by selection. Also it was found that conservation of synteny was related to similarity in expression of yeast gene pairs.

Acknowledgements

Firstly, I wish to thank my supervisor, Laurence Hurst, for his help and patience in getting me through these past few years. I also thank everyone who has worked with me over the years, in particular, Csaba Pal (Chapters 7 & 8) and Martin Lercher (Chapter 3) who contributed to several of the papers presented in this thesis. Araxi Urrutia was a great help over the two years I have worked with her. Especially her help in teaching me how to write English and proof reading various documents I have attempted to write. Thanks also to my internal Mike Mogie. My source of funding was from a University of Bath studentship. I would also like to thank everyone else who has put up with me for the last three years; Kenton, Mark, Danny, Jill, Hamilton, Mother, Daddy, Patricia, Frances, Margaret, and of course Galaxy (who maybe got more attention than she deserved).

Contents

Introduction

Evolution of genes

The question of how genes evolve is central to the field of evolutionary biology as most evolutionary change occurs primarily at the gene level. However, the rate of substitutions does not appear to be constant for all genes. Different genes evolve at different rates. For example, genes involved in the immune response seem to evolve faster possibly due to antagonistic co-evolution (Hughes and Nei, 1988; Hurst and Smith, 1999).

One of the major questions asked in molecular evolution is why is there variation in rates of evolution? By looking at the rates and patterns of gene evolution it is possible to answer questions concerning the processes that cause these differences (Li, 1997). The two main influences on gene evolution are selection, that is how a change at the level of the sequence of a gene affects the overall fitness of an individual, and genetic drift, where random influences, which allow alleles that are not as fit to increase in a population and superior alleles to be lost from the population (Gillespie, 1991). By looking at the rate of evolution of genes it is possible to determine which of these two influences is the dominant effect driving the evolution of a species or even a particular gene. Most studies, that attempt to explain why genes evolve at a certain rate, compare the rate of evolution of a gene with the parameters that may affect the rate. For example, collating a dataset of the evolutionary rates of orthologous genes and comparing

this to either a factor known to be under selective pressure, such as expression patterns, or one that is thought to be due to a bias in mutation rates.

The ability to answer these questions has increased due to the vast amount of gene data that has recently become available, both sequence and expression information. Early work was hampered by small datasets as well as lack of computational power, for example alignments had to be done by hand. Both of these problems have been overcome because of the genome sequencing projects and the expansion of the field of bioinformatics. This has allowed much larger datasets to be examined, both in the number of genes and in the number of parameters to which evolutionary rates can be compared. Due to these improvements more detailed analyses of the factors that could affect the evolution of genes can be made, both in terms of which factors affect gene evolution and also the interactions between these factors.

Measurement of rates of evolution

In order to determine which processes affect rates of evolution there needs to be an accurate method for measuring or estimating these rates. The rate of evolution of genes is not a measure of rate of change over time, in reality what is being measured is the proportion of sites that have changed in the time since divergence of two nucleotide sequences. All of the methods that will be described in this introduction start with a pair of aligned nucleotide sequences. The algorithms begin by counting the number of sites in the sequences and then count the number of nucleotide differences between the sequences. One difficulty is in

determining whether there have been any multiple hits i.e. more than one nucleotide substitution at any one site. The difference between the methods is primarily in the way multiple hits are calculated.

The development of methods used to measure rates of evolution.

The first method to measure rates of evolution was developed by Jukes and Cantor (JC) (1969). This is a simple one-parameter model whereby all mutations are considered to occur at equal rates. However, this is an unrealistic model for several reasons, one of these reasons being that transitions occur more frequently than transversions (Li, 1997). An alternative model was proposed by Kimura, which is a two-parameter model (Kimura-2-parameter; K2P) (Kimura, 1980). This method takes into account the difference in mutation rates between transitions and transversions. Both of these models, JC and K2P, can be used to calculate K, which is defined as the number of substitutions per site since the point of divergence between the two sequences. This measurement is appropriate if the sequence is non-coding, but when the sequence encodes protein a distinction needs to be made between synonymous and non-synonymous changes. Synonymous changes are silent mutations where a change at the nucleotide level does not change the protein sequence. Conversely a non-synonymous change is one that alters the protein sequence. It is important to make this distinction, because the rate of change at each of these sites will be affected differently by the forces that affect evolutionary rates. Thus, two values are produced when measuring rates of

evolution in protein coding sequences, Ka and Ks. Ka (or dN) is defined as the number of non-synonymous substitutions per non-synonymous site whereas Ks (or dS) is defined as the number of synonymous substitutions per synonymous site.

An added complication exists due to the structure of the genetic code whereby some sites can undergo both a synonymous or non-synonymous change. A method, which takes into account the problems associated with calculating the evolutionary rates of protein coding sequences, was developed simultaneously by Li (Li, 1993) and by Pamilo and Bianchi (Pamilo and Bianchi, 1993). Both methods divide the sequence into three types of site, non-degenerate, 2-fold degenerate and 4-fold degenerate sites, where the degeneracy of a site is determined by the number of alternative nucleotides that encode for the same amino acid at that site. For example, a 4-fold degenerate site encodes for the same amino acid no matter which nucleotide is present, whereas at a 2-fold site two alternative nucleotides encode for the same amino acid. Ka can then be calculated from the number of substitutions at non-degenerate and 2-fold degenerate and Ks from the number of substitutions at 2-fold and 4-fold degenerate sites. K2P is used to calculate the number of multiple hits. All of these methods are termed approximate methods in that they incorporate *ad hoc* assumptions of sequence evolution (Yang and Nielsen, 2000). A recently developed maximum likelihood method is considered to have solved these problems (Goldman and Yang, 1994). This method uses codon frequencies to determine the parameters in the model. An approximate version of this method was also developed (YN00) which gives very similar results (Yang and Nielsen, 2000). The work outlined in this thesis will use

all of these methods to calculate rates of evolution although the majority of the work uses the methods described in Li (1993) and Pamilo and Bianchi (1993).

Variation in rates of evolution

The rate at which genes evolve varies considerably between genes. Graur and Li (2000) collated a dataset of 47 orthologous protein-coding genes from human and mouse. It was found that there was substantial variation in the rates of protein evolution and mutation. Ka was found to vary from 0 to $3.1x10^{-9}$ non-synonymous substitutions per site. The expectation will be that the variation in protein coding sequences is due to variation in selective constraint. This was confirmed in a later study where it was found that there is a strong negative correlation between Ka and tissue expression breadth, which is the number of tissues in which a gene is expressed (Duret and Mouchiroud, 2000). This is consistent with the proposal that variation in selective constraint in the form of expression breadth causes variation in rates of protein evolution. The expectation for Ks, according to the neutral theory of evolution, is that the silent substitution rate should equal the mutation rate (Kimura, 1983). Therefore, if the genomic mutation rate is constant, there should be no variation in Ks. However, there is a great deal of evidence showing variation in Ks.

Evidence that there is variation in mutation rates

Wolfe et. al. (1989) studied the variation in silent substitutions in mouse and rat orthologues. It was found that there was significant variation in K4 (substitution

rate at four-fold degenerate sites) calculated from 23 mouse-rat orthologues (Wolfe et al., 1989). It was suggested that this variation was regional in nature, although the dataset was small. A more thorough study using 363 mouse-rat orthologues confirmed that there was variation in substitution rates (Wolfe and Sharp, 1993). This showed the extent of variation in Ks in comparison to Ka and pointed to the effect possibly being regional due to GC% (proportion of guanine and cytosine nucleotides) variation (Wolfe and Sharp, 1993). Koop (1995) also investigated the regional variation in mutation rates, however only 8 genes were used of which only 3 were in the same region. Similarly Mouchiroud et. al. (1995) found variation in Ks, but demonstrated that the variation seemed to be gene specific suggesting that silent sites were under selection. However, this study used a less accurate method to calculate Ks (LWL (Li et al., 1985)), which is biased by GC%. Therefore the repeatability in Ks could simply be due to conservation of nucleotide content between orthologous genes. Most of the studies that showed variation in rates of silent substitutions suggested that this variation was regional in nature (Wolfe and Sharp, 1993; Koop, 1995). This effect was confirmed by Matassi et. al. (1994; 1999) who compiled a dataset of human-mouse orthologues and demonstrated that linked genes had similar Ks. This suggests that the cause of the variation in Ks is also regional in nature.

Evidence that the mutation rate is constant over the entire genome

Recently Kumar and Subramanian (2002) suggested that mutation rates were constant over the genome and that any variation observed in estimates of Ks or K4 was due to a difference in substitution patterns between orthologous genes. This difference was assumed to be due to a shift in the nucleotide composition of a gene caused by the movement of a gene to a region with a different background GC% content. They compiled a large dataset of 5,669 nuclear genes from a total of 326 species to examine the extent of variation in mutation rates within and between lineages. From this dataset they removed any orthologues that showed a disparity in the pattern of substitutions, thus preferentially removing genes with a high GC%. This led to the conclusion that there was no variation in the mutation rate over the genome and that gene location does not affect the mutation rate (Kumar and Subramanian, 2002). This study has been questioned by Smith et. al. (2002) on the basis that recent findings have found that nucleotide composition is non-stationary (Duret et al., 2002). Therefore, the rationale for the removal of sequences that show a disparity in the substitution process is questionable.-Assuming this is the case the conclusion that there is regional variation in the silent substitution rate seems relatively robust.

Theoretical explanations for why rates of evolution could differ

There are many explanations for why there is variation in rates of evolution, in particular why the variation could be regional. In order to fully understand the process of gene evolution, explanations for the variation in Ka as well as Ks need to be considered. First why is there variation in Ks? The reasons given for such variation mostly involve some form of mutational bias, but selection and biased gene conversion have also been considered in order to explain the variation in nucleotide composition. Second, the explanations for variation in Ka will be considered. Since Ka is a measure of the rate of protein evolution, the explanations for variation in protein evolution examine possible causes for variation in selection. Either selection pressure, the function of the protein, determines the rates of evolution, or alternatively selective efficiency, since theoretical work suggests that selection may not be as effective in some genomic regions compared to other regions.

Explanations for variation in mutation rates

Four explanations for the variation in Ks will be considered. The first is male mutation bias, which was originally described in the 1930s (Haldane, 1935). An alternative explanation for variation in mutation rates due to mutation bias is the replication-timing model (Filipski, 1987). This model predicts that different regions of the genome, because they are replicated at different times, will have

disparate mutation rates due to variation in the availability of dNTPs during the cell cycle. It has also been found that recombination itself induces mutations. Thus, since recombination is variable around the genome, mutation rates could too be displaying the same pattern. However, a fourth explanation for the variation in Ks is that the variation is directly caused by variation in the nucleotide composition of the genome. Three theories, in this context, which attempt to explain mutation variation in terms of variation in GC% will be discussed; mutation bias, selection and biased gene conversion.

1) Male mutation bias

One of the important clues to understanding the variation in Ks comes from the finding that males seem to have a higher mutation rate than females (Crow, 1997). This was first demonstrated by Haldane when he showed that a much greater proportion of mutations leading to X-linked haemophila originated in men compared with women (Haldane, 1935; Haldane, 1947). This early work was confirmed by a more recent study carried out by Rosendaal et. al. (1990). Similarly, the same conclusions have been found for many autosomal dominant genetic diseases (see Table 1. for details). The cause of this male biased mutation rate was thought to be the larger number of cell divisions required for the production of male gametes (for example, sperm in mammals). For example, the sperm of a 28-year-old male has already undergone 380 mitoses (Risch et al., 1987; Plas et al., 2000), compared to oocytes that only undergo 33 mitoses in females (Chang et al., 1994).

Evidence for a male bias in mutation rates

Two lines of evidence suggest that there is a male bias in mutation rates. The first is that mutation rates increase with paternal age (Lian et al., 1986; Risch et al., 1987). The second is based on the possibility that such a male biased mutation rate would lead to different mutation rates on the X, Y and autosomes (Miyata et al., 1987). Since the Y chromosome is only found in males it should show a much higher mutation rate than the X chromosome. Autosomes, spending equal time in both sexes, are predicted to have an intermediate mutation rate. Estimating the male mutation bias, α (male mutation rate/ female mutation rate) using the variation in rates between the X, Y and autosomes has been calculated for many different species. Estimates from primates vary from 4.2 (Chang et al., 1996) to 6.0 (Shimmin et al., 1993). An average value of 5.06 was calculated using introns from three primate genes, ZFX/ZFY, AMGX/AMGY and SMCX/SMCY (Huang et al., 1997). When calculating α from birds the mutation rates on the Z and W chromosomes are used, where the females are the heterogametic sex (ZW). In this case the W chromosome should have a higher mutation rate if there is a male biased mutation rate. Comparing the substitution rate at silent sites (Ks) as well as in introns on the Z and W chromosomes of birds α was estimated at 3.9-6.5. This confirmed that there was a significant increase in the mutation rate in males (Ellegren and Fridolfsson, 1997).

The evidence for male mutation bias is weak

There is alternative evidence however, which suggests that the evidence for male mutation bias is weak or non-existent. Evidence from dominant genetic diseases, for example (Table 1.), shows that the calculation of α does not always

give the same value, and many do not show $\alpha > 1$. A value for α greater than one is expected if there is male mutation bias. In addition, the increase in mutation rate with paternal age does not explain the increase in the occurrence of paternally derived mutations (Risch et al., 1987; Tiemann-Boege et al., 2002). This was tested by looking directly at mutations in fibroblast growth factor receptor 3 found in sperm by using PCR. There was an increase in male mutation rates, but not an exponential increase as expected from the increase in disease occurrence with paternal age (Tiemann-Boege et al., 2002).

Recent estimates of α calculated in primates have also been surprisingly low ($\alpha = 1.66$) (Bohossian et al., 2000). Makova and Li (2002) have refuted these findings as Bohossian et. al. (2000) made no correction for polymorphism levels between closely related species. They also assumed a particular time of translocation of their X segment to the Y. If the translocation occurred at a later time the segment would have been X linked for a longer period of time and misleading estimates of α would have been estimated. For this reason Makova and Li (2002) re-estimated α from a Y/A comparison rather than a Y/X comparison using more distantly related species ($\alpha = 5.25$).

While the challenge to the estimate for alpha in humans seems to have been rebutted problems remain. Most notably, at least in rodents, the estimate depends on whether we compare X with Y or X with autosomes. Smith and Hurst (1999a), following McVean and Hurst (1997) show that X-Y comparisons reveal alpha to be around 2, which is consistent with germ line replication. However, X-A comparisons suggest that nearly all mutations are male derived. This heterogeneity

is itself evidence against the replication timing model (described below), as all comparisons within a given species pair should give the same figures, but its explanation in unknown.

2) Replication timing

The variation in Ks between autosomal genes (Wolfe and Sharp, 1993), as described earlier cannot be completely explained by the male mutation bias hypothesis. This is simply because the male mutation bias theory predicts that autosomal genes should have a constant mutation rates, assuming all other things are equal. An alternative explanation for the variation in Ks is that the accuracy of DNA replication may depend on dNTP availability in the cell [Wolfe, 1989 #6556]. It appears that dNTPs are a limited resource and it has been shown in cultured mouse fibroblast cells that there are fluctuations in dNTP availability during the cell cycle (McCormick et al., 1983). It has also been demonstrated that variation in the dNTP pool in the cell causes changes in the fidelity of DNA replication (Bebenek et al., 1992). Since DNA replication occurs in different genomic regions at different times during the cell cycle (Tadokoro et al., 2002), there is therefore the possibility of regional variation in mutation rates. A model of the replication process developed by Wolfe (1991) suggested that variation in dNTP availability could cause variation in mutation rates. However, this model made many assumptions which are not valid; such as non-independence of mutations and lack of bias in mismatch repair.

Variation in dNTP pools is also used as an explanation for the existence of GC% variation around the genome. For example, early replicating DNA was

thought to be GC-rich and late replicating DNA GC-poor. The same pattern was also reflected in the G-banding pattern found in mammalian chromosomes (Holmquist et al., 1982). However, the timing of replication and nucleotide composition of a gene does not seem to be related (Eyre-Walker, 1992a). Therefore, this explanation for mutational variation is weak and as such some other cause needs to be considered.

3) Recombination induced mutations

Recently, the notion that recombination might be mutagenic has been attracting increasing attention. Early work suggested that the mutation rate during meiosis was higher than the spontaneous mitotic rate, and many of the mutations appear to be associated with nearby crossover events (Magni and von Borstel, 1962; Magni, 1964; Esposito and Bruschi, 1993). A mechanistic basis for this effect may be faulty repair of the double strand breaks that initiate recombination. Such an effect has been demonstrated experimentally in mitotic recombination in yeast (Strathern et al., 1995; Rattray et al., 2001). A comparable mechanism may also underlie somatic hypermutation in mammals (Papavasiliou and Schatz, 2000). A similar effect has been postulated in other fungi (Yeadon and Catcheside, 1999).

The possibility that recombination is intrinsically mutagenic is consistent with observations on the pseudoautosomal region in humans, which recombines during the pairing of X and Y chromosomes in dividing male cells, leading to a very high rate of recombination. The same region is not only known to be highly polymorphic (Schiebel et al., 2000), but also shows elevated rates of synonymous substitution (Perry and Ashworth, 1999). Indeed, in the mouse-rat comparison

pseudo-autosomal genes also show unusually high synonymous substitution rates (L. Hurst, Unpublished data). In a genome wide comparison of human to rodent sequences it is claimed that Ks covaries with the recombination rate (Lercher and Hurst, 2002), but as the ancestral recombination rate of any sequence is hard to know over such a broad time span, this finding requires confirmation from more closely related species. A mutagenic effect of recombination has also been suggested as an explanation for a putative relationship between codon usage bias and recombination rates in flies (Marais et al., 2001).

4) Regional variation in GC%

Isochores

Evidence that variation in GC% could lead to variation in substitution rates is that the variation in GC% appears to be regional in nature. The first evidence for regionality in GC% was the finding that genomic DNA separated into several DNA bands using caesium chloride gradients (Theirry et al., 1976). Each band of DNA in the gradient had a different GC content. Along with the finding that the GC% of a gene is correlated with the GC% of the surrounding genomic DNA, this banding pattern led to the proposal of the isochore structure of mammalian genomes where different genomic regions have varying GC% (Bernardi et al., 1985).

There are two different definitions that have been used for isochore structure. One is the strict definition where there are identifiable boundaries between the different isochores, the other where there is simply local similarity in

GC with no definite boundaries (Eyre-Walker and Hurst, 2001; International Human Genome Sequencing Consortium, 2001). There is very little evidence for a strict definition of isochores. For example, there is no evidence for isochore boundaries on chromosomes 21 and 22 in humans (Haring and Kypr, 2001). In addition, a complete survey of the draft version of the human genome sequence failed to find any evidence of isochores in this "strict" sense (International Human Genome Sequencing Consortium, 2001). Bernardi, however, disputes that isochores, when discovered, were defined in this manner (Bernardi, 2001).

Nevertheless, there is strong evidence for closely linked regions of the genome to have similar GC%, the second definition of isochores. For example it has been shown that neighbouring regions of the genome have greater similarity in GC% than expected by chance (Matassi et al., 1994).

The question is what causes the local similarity in GC%, selection, mutation or biased gene conversion. It is hoped that by investigating the relationship between Ks and GC% it may be possible to determine what caused the regional variation in GC%.

Thus, Ks can be used to determine the cause of the regional variation in GC%. However, a correlation between Ks and GC% could also cause local similarity in Ks simply because there is strong local similarity in GC%. Importantly, the reason Ks can be used to explore the explanation for the regional variation in GC% is that different hypotheses predict different correlations between Ks and GC% (Nagylaki, 1983; Gu and Li, 1994). The three theories, selection,

mutation bias and biased gene conversion, which predict the cause for the variation in genomic GC%, will be now be considered in detail.

i) Selection

The selectionist hypothesis for the evolution of isochores suggested by Bernardi (1989) is that the GC rich isochores act as staples holding the DNA together. This is because the G-C bond linking the two strands of DNA requires 3 hydrogen bonds whereas A-T bond requires just two (Alberts et al., 1994). The suggestion is that in species with high body temperatures such as mammals there is strong selection for regions with high GC in order to keep the DNA strands intact (Bernardi, 2000). This theory predicts that there is a correlation between GC% and growth temperature in a wide range of organisms, and that isochores are only found in animals that are warm–blooded (Bernardi, 2000). These predictions have been questioned. No correlation was found between genomic GC% and growth temperature of bacteria suggesting selection is not acting to increase the GC% of bacteria that live in hot environments [Galtier, 1997 #9685](Hurst and Merchant, 2001). More recently the absence of any correlation has also been shown in vertebrates. Evidence also suggests that genomic variation in GC% is found in crocodiles and turtles, which are cold blooded and therefore, according to Bernardi's hypothesis, should not show evidence of genomic variation in GC% (Hughes et al., 1999).

ii) Mutation

There are several different proposals for how mutation bias could cause the variation in mutation rates within the genome. Most of these proposals rely on the variation in nucleotide composition as evidence for the model. A distinction between the mutational bias models can be made in how Ks relates to GC%. For example, an inverted U/V shaped relationship is evidence for the dNTP or nucleotide mis-incorporation model (Gu and Li, 1994), whereas a positive correlation could be expected if the bias is due to the high mutation rate of CpG dinucleotides.

The inverted U/V shaped relationship between Ks and GC% was found by Wolfe et. al. (1989) (as well as by Bulmer et. al. (1991)) and was used as evidence for the replication timing model. This relationship between Ks and GC% is theoretically possible but is highly sensitive to the efficiency of proofreading during replication (Wolfe, 1991). Nevertheless, Bernardi disputed Wolfe's finding of an inverted U shaped distribution claiming that the sample size was too small (Bernardi et al., 1993) but see (Wolfe and Sharp, 1993). Possibly more importantly, Mouchiroud also claims that the same finding was due to a methodological artefact (Mouchiroud et al., 1995).

Surprisingly, the use of a maximum likelihood method developed by Goldman and Yang (Goldman and Yang, 1994), showed a strong positive correlation (Smith and Hurst, 1999b; Bielawski et al., 2000). A possible explanation for this correlation is the effect of CpG deamination. CpG to TpG mutations are expected to occur at much higher frequencies than other mutations

such that at high GC the mutation rate would be higher (Giannelli et al., 1999). Piganeau et. al. (2002) developed a mutational model incorporating this type of mutation, which suggested that such a process could lead to the positive correlation between Ks and GC%. Fryxell and Zuckerkandl (2000) also developed an interesting theory of how cytosine deamination could lead to the formation of isochores based on the premise that AT rich sequences are more likely to open by "DNA breathing" and hence become more susceptible to mutations towards AT. This would obviously lead to a negative feedback loop, which could result in AT rich sequences becoming more AT rich and GC rich sequences becoming more GC rich (Fryxell and Zuckerkandl, 2000).

The mutational model is an attractive explanation for the existence of the regionality in GC%, not least because it does not require us to suppose that a silent GC<->AT change in a sea of non-coding DNA is associated with selective deaths. However, theoretical models suggest that the effect is parameter sensitive (Wolfe, 1991; Eyre-Walker, 1992b). Since different models predict different relationships between Ks and GC% it seems that any new correlation can be explained by some form of mutational bias.

iii) Biased gene conversion

The third proposal to explain the regional variation in GC is biased gene conversion (BGC) (Holmquist, 1992; Eyre-Walker, 1993; Eyre-Walker, 1999; Galtier et al., 2001). This process occurs when a recombination event happens at a polymorphic site, at which point the mismatch repair system is recruited to repair the mismatch. A conversion event then occurs as one allele is "repaired". For

example, if there is a G-T mismatch between the two strands of DNA, either the T will be replaced with a C, or the G with an A (Galtier et al., 2001). This process increases the probability of one of the alternate bases reaching fixation in the population. The result is a relationship between the recombination rate and the substitution rate, as mutations will reach fixation faster in regions of increased recombination. However, there is evidence that there is a bias in mismatch repair towards GC (Brown and Jiricny, 1989). If this were so, the T, as described in the above example, will be more often replaced with a C than vice versa.

A consequence of such biased gene conversion would be that in regions of high recombination, GC% could be predicted to increase. Since recombination varies in a regional manner (i.e. there are hot spots and cold spots of recombination (Broman et al., 1998)) GC% could also become regional. This hypothesis predicts a decline in Ks as GC% reaches 100%, but in addition predicts a positive correlation between GC% and recombination (Galtier et al., 2001). Several studies have found such a correlation (Eyre-Walker, 1992a; Fullerton et al., 2001; Birdsell, 2002). Fullerton et. al (2001), for example, compared the GC% calculated from intron sequences and the recombination rate which was measured as cM/Mb. They found a significant positive correlation between GC% and recombination rate, which is predicted by BGC. On a finer scale, Eisenbarth et al. (2000) showed that a large increase in the extent of linkage disequilibrium was found at a boundary between a GC rich and GC poor genomic region.

The above sections illustrate that there is an ongoing debate regarding the cause of the variation in the silent substitution rate. However, it should be possible

to determine what is causing the variation in Ks by comparing rates of evolution to nucleotide composition, recombination and other indicators of mutation bias.

Explanations for the variation in rates of protein evolution

Several explanations for the variation in the silent rate of evolution have been considered in the previous sections. However, different forces need to be considered when analysing the variation in the rate of protein evolution. This requirement arises because amino acid substitutions are more likely to be influenced by selective pressures. Thus, any variation in rates of protein evolution is likely to be due to variation in selection in addition to underlying differences in the mutation rate. This consequence could either be due to variation in the flexibility of the amino acids required for function, i.e. some proteins could work just as well if some of the amino acids are replaced with ones with similar attributes, or alternatively the consequence could arise from variation in the efficiency of selection.

1) Variation in selective efficiency

The efficiency of selection can simply be described as the ability of selection to act on slightly deleterious mutations. These mutations are defined as possessing a selection coefficient $s \approx 1/2Ne$. Since the selection coefficient is dependent on Ne (the effective population size) and recombination effectively increases the effective population size, regions with increased recombination will be subject to more efficient selection (Hill and Robertson, 1966). This effect is

described in different ways including the Hill-Robertson effect, background selection (Charlesworth, 1994) and genetic hitchhiking (Kliman and Hey, 1993). However, these processes only work on slightly deleterious mutations. Therefore, these effects predict that most of the variation in rates of protein evolution is due to variation in the proportion of slightly deleterious mutations which reach fixation. This requires that there is a substantial number of potential mutations which do not adversely affect the phenotype of an individual.

2) Variation in selective pressure

Studies suggest that the function of a gene has a strong influence on the rate of gene evolution. For example, immune genes tend to evolve fast (Hughes and Nei, 1988; Hurst and Smith, 1999), whereas neuronal genes evolve slowly (Kuma et al., 1995). Similarly, the extent to which sub-parts of the protein are functional is typically assumed to explain variation between proteins (Kimura, 1983). It is typical for example to find that active centres are highly conserved. Indeed it is owing to this that we can attempt to make guesses at the function of a protein form the amino acid sequence.

A comparable model was tested by asking whether more dispensable genes might evolve faster than less dispensable ones (Hirsh and Fraser, 2001). The premise of this test is that in essential genes a higher proportion of mutations will be eliminated by selection than in more dispensable genes in which we expect more effectively neutral ($s<<1/2Ne$) or slightly deleterious mutations. While initial reports claimed no effect of dispensability on the rate of evolution (Hurst and

Smith, 1999), Hirsh and Fraser (2001) using growth rate data of yeast with given gene knockouts found a negative correlation between dispensability and rate of evolution as predicted (Wilson et al., 1977).

The above result however leaves open the issue of what determines dispensability. Possibly one interacting factor is the expression pattern of the gene (Pal pers. comm.). Importantly, Duret and Mouchiroud (2000) demonstrated that broadly expressed genes have lower rates of protein evolution. In addition, Pal et. al. (2001) showed that, in yeast, genes which have high expression levels, based on mRNA levels in the cell, evolve slowly. This evidence points to variation in selective pressure, measured, for example, in terms of expression patterns, as being an important cause of variation in rates of protein evolution.

Emerging evidence that genes of similar function and expression are clustered in the genome

In order for selective pressure to adequately explain regional patterns in rates of protein evolution, selection needs to vary depending on the genomic location of the gene. In practise this could mean that genes with similar function or expression cluster in the genome, such that genes with similar selective pressure are clustered.

There is a great deal of evidence suggesting that clustering of genes with similar function or expression occurs in the genome, and the evidence is rapidly increasing. In bacteria there are well known examples of gene clusters involved in the same

pathway. These clusters, termed operons, are often involved in the same biochemical pathway and are also co-transcribed (Alberts et al., 1994). In eukaryotes there are several examples of gene clusters with similar functions, the Hox cluster being the most well known (Holland and Garciafernandez, 1996; Brooke et al., 1998). Recently evidence of operons in eukaryotes has also been uncovered, albeit in just one species, *Caenorhabditis elegans* (Zorio et al., 1994; Blumenthal et al., 2002; for review see Nimmo and Woollard, 2002).

Despite the lack of evidence for widespread use of operons, there is increasing evidence in eukaryotes that expression pattern and physical position are related variables. Using yeast microarray data, Cohen et. al. (2000) were able to show that pairs of adjacent genes had greater similarity in expression patterns than expected compared to non-adjacent gene pairs. By integrating the human genome map with expression data derived from SAGE data (Serial Analysis of Gene Expression), Caron et. al. (2001) showed that there was strong clustering of highly expressed genes. A more recent study in Drosophila, again using microarray data, showed that groups of adjacent genes with similar expression patterns were found frequently in the real dataset but rarely in any randomised dataset (Spellman and Rubin, 2002). These results are strong evidence that linked genes have similar expression patterns or similar function.

Due to the clustering of genes with similar expression there is a strong possibility that linked genes would have similar rates of protein evolution.

The aims of this thesis

The aim of this thesis is to address the causes of variation in evolutionary rates, particularly why there is regional variation in Ks and whether the variation in Ka is also regional. This series of research projects started at a time when there was no strong evidence for regional variation in mutation rates and no evidence that linked genes had similar rates of protein evolution. In the first results chapter of this thesis (Chapter 2) regional variation of rates of evolution in a dataset of mouse-rat orthologues was demonstrated. This illustrated for the first time that linked genes did indeed have similar rates of protein evolution (Williams and Hurst, 2000). Chapter 3 compared the same mouse-rat dataset to a human-rodent dataset looking at local similarity in rates of evolution using a more sophisticated statistical method. This showed that there was significant chromosomal heterogeneity in mutation rates (Lercher et al., 2001).

The question of whether the variation in Ks can be explained by selection was

examined in Chapter 4. In this chapter the evidence for gene specific Ks was questioned through the use of a variety of different methods to calculate silent rates of evolution. The extent of gene specificity in Ks was shown to depend on the method employed to calculate substitution rates (Williams and Hurst, 2002b). The study also analysed the relationship between Ks and GC%, which is central to the debate on Ks variation. The same issue was studied in Chapter 5 when the approximate version of the maximum likelihood method, YN00, was used to

calculate rates of evolution. The study confirmed previous findings that there is a positive relationship between Ks and GC% (Hurst and Williams, 2000).

Chapter 6 examines variation in expression patterns in the mouse genome and asks whether the variation in expression breadth can explain the regional variation in rates of evolution. It was shown that linked genes have similar expression breadth and that this pattern explains some of the local similarity in rates of protein evolution (Williams and Hurst, 2002a). In order to answer the same problem, Chapter 7 shows a study of intragenic variation in rates of evolution. This intragenic variation was used to show that sequences with the same expression pattern but different functions have similar rates of evolution (Williams et al., 2000).

Finally, given the accummulating evidence that gene location is not random I ask whether we can find evidence that selection prefers certain gene arrangements above others. Chapter 8 therefore examines whether the clustering of genes with similar expression patterns is conserved between species. It was shown that similarity in expression patterns is a strong indicator for conservation of synteny (Hurst et al., 2002).

In summary this thesis has examined the variation in rates of evolution of protein coding genes. Evidence is published for the first time that linked genes have similar rates of protein evolution. The causes for this observation are thoroughly examined. The implications of these findings for our general understanding of the forces that affect evolution are presented in detail in the final discussion (Chapter 9).

TABLE 1.

Estimates of the ratio (α) of the number of point mutations of paternal origin to those of maternal origin leading to dominant autosomal disorders of humans. N is the number of informative independent mutations.

Disease	Gene	mutation type	α	N	Refs
Multiple endocrine neoplasia type 2B (MEN 2B)§	*RET*	point	∞	25	(Carlson et al., 1994)
MEN 2A	*RET*	point	∞	10	(Schuffenecker et al., 1997)
Hirschsprung disease	*RET*	point	0	3	(Yin et al., 1996)
Achondroplasia	*FGFR3*	point	∞	40	(Tiemann-Boege et al., 2002)
Pfeiffer Syndrome	*FGFR2*	point	∞	11	(Nagase et al., 1998)
Crouzon Syndrome	*FGFR2*	point	∞	11	(Nagase et al., 1998)
Apert syndrome	*FGFR2*	point	∞	57	(Moloney et al., 1996)
Neurofibromatosis 2	*NF2*	point	1.3	23	(Kluwe et al., 2000)
Neurofibromatosis type 1	*NF1*	point?Ý	4.5#	11	(Lazaro et al., 1996)
Hamartoma syndrome tuberous sclerosis	*TSC2*	point	0.66	5	(Roberts et al., 2002)
Von Hippel-Lindau Disease	*VHL*	point	1.3	7	(Richards et al., 1995)
Retinoblastoma	*RB1*	not large	8.5	38	(Dryja et al., 1997)

deletions¶

§ NB maternally derived mutations have now been described

Ý Probably point mutations but may be small deletions

Other reports show a higher male bias ($\alpha=f$, N=10;α=6, N=14) but the sort of mutations are unknown.

¶ May be either point mutations or small deletions

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D.: Molecular Biology of the Cell. Garland Publishing, Inc., New York and London, 1994.

Alvarez-Valin, F., Jabbari, K. and Bernardi, G.: Synonymous and nonsynonymous substitutions in mammalian genes: Intragenic correlations. Journal of Molecular Evolution 46 (1998) 37-44.

Bebenek, K., Roberts, J.D. and Kunkel, T.A.: The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. Journal of Biological Chemistry 267 (1992) 3589-96.

Bernardi, G.: The Isochore organization of the human genome. Annual Review of Genetics 23 (1989) 637-661.

Bernardi, G.: Isochores and the evolutionary genomics of vertebrates. Gene 241 (2000) 3-17.

Bernardi, G.: Misunderstandings about isochores. Part 1. Gene 276 (2001) 3-13.

Bernardi, G., Mouchiroud, D. and Gautier, C.: Silent substitutions in mammalian genomes and their evolutionary implications. Journal of Molecular Evolution 37 (1993) 583-589.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunierrotival, M. and Rodier, F.: The Mosaic Genome of Warm-Blooded Vertebrates. Science 228 (1985) 953-958.

Bielawski, J.P., Dunn, K.A. and Yang, Z.H.: Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. Genetics 156 (2000) 1299-1308.

Birdsell, J.A.: Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Molecular Biology and Evolution 19 (2002) 1181-1197.

Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M. and Kim, S.K.: A global analysis of Caenorhabditis elegans operons. Nature 417 (2002) 851-854.

Bohossian, H.B., Skaletsky, H. and Page, D.C.: Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature 406 (2000) 622-625.

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. and Weber, J.L.: Comprehensive human genetic maps: individual and sex-specific variation in recombination. American Journal of Human Genetics 63 (1998) 861-9.

Brooke, N.M., Garcia-Fernandez, J. and Holland, P.W.H.: The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. Nature 392 (1998) 920-922.

Brown, T.C. and Jiricny, J.: Repair of base-base mismatches in simian and human cells. Genome 31 (1989) 578-83.

Bulmer, M., Wolfe, K.H. and Sharp, P.M.: Synonymous nucleotide substitution rates in mammalian genes - implications for the molecular clock and the relationship of mammalian orders. Proceedings of the National Academy of Sciences of the United States of America 88 (1991) 5974-5978.

Carlson, K.M., Bracamontes, J., Jackson, C.E., Clark, R., Lacroix, A., Wells, S.A. and Goodfellow, P.J.: Parent-of-origin effects in multiple endocrine neoplasia type 2b. American Journal of Human Genetics 55 (1994) 1076-1082.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heisterkamp, S., van Kampen, A. and Versteeg, R.: The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291 (2001) 1289-1292.

Casane, D., Boissinot, S., Chang, B.H.J., Shimmin, L.C. and Li, W.H.: Mutation pattern variation among regions of the primate genome. Journal of Molecular Evolution 45 (1997) 216-226.

Chang, B.H.J., Hewett-Emmett, D. and Li, W.-H.: Male-to-female ratios of mutation-rate in higher primates estimated from intron sequences. Zoological Studies 35 (1996) 36-48.

Chang, B.H.J., Shimmin, L.C., Shyue, S.K., Hewett-Emmett, D. and Li, W.-H.: Weak male-driven molecular evolution in rodents. Proceedings of the National Academy of Sciences of the United States of America 91 (1994) 827-831.

Charlesworth, B.: The effect of background selection against deleterious mutations on weakly selected, linked variants. Genetical Research 63 (1994) 213-227.

Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M.: A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nature Genetics 26 (2000) 183-6.

Crow, J.F.: The high spontaneous mutation rate: Is it a health risk? Proceedings of the National Academy of Sciences of the United States of America 94 (1997) 8380-8386.

Dryja, T.P., Morrow, J.F. and Rapaport, J.M.: Quantification of the paternal allele bias for new germline mutations in the retinoblastoma gene. Human Genetics 100 (1997) 446-449.

Duret, L. and Mouchiroud, D.: Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Molecular Biology and Evolution 17 (2000) 68-74.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D. and Galtier, N.: Vanishing GC-rich isochores in mammalian genomes. In: Bernardi, G. (Ed.), Molecular evolution: evolution, genomics and bioinformatics, Sorrento(Naples), 2002, pp. 65.

Eisenbarth, I., Vogel, G., Krone, W., Vogel, W. and Assum, G.: An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. American Journal of Human Genetics 67 (2000) 873-880.

Ellegren, H. and Fridolfsson, A.K.: Male-driven evolution of DNA sequences in birds. Nature Genetics 17 (1997) 182-184.

Esposito, M.S. and Bruschi, C.V.: Diploid yeast cells yield homozygous spontaneous mutations. Current Genetics

23 (1993) 430-434.

Eyre-Walker, A.: Evidence that both g+c rich and g+c poor isochores are replicated early and late in the cell-cycle. Nucleic Acids Research 20 (1992a) 1497-1501.

Eyre-Walker, A.: The role of dna-replication and isochores in generating mutation and silent substitution rate variance in mammals. Genetical Research 60 (1992b) 61-67.

Eyre-Walker, A.: Recombination and mammalian genome evolution. Proceedings of the Royal Society of London Series B 252 (1993) 237-243.

Eyre-Walker, A.: Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. Genetics 152 (1999) 675-683.

Eyre-Walker, A. and Hurst, L.D.: The evolution of isochores. Nature Reviews Genetics 2 (2001) 549-555.

Filipski, J.: Correlation between Molecular Clock Ticking, Codon Usage, Fidelity of DNA-Repair, Chromosome-Banding and Chromatin Compactness in Germline Cells. Febs Letters 217 (1987) 184-186.

Fryxell, K.J. and Zuckerkandl, E.: Cytosine Deamination plays a primary rold in the evolution of mammalian Isochores. Molecular Biology and Evolution 17 (2000) 1371-1383.

Fullerton, S.M., Carvalho, A.B. and Clark, A.G.: Local rates of recombination are positively correlated with GC content in the human genome. Molecular Biology and Evolution 18 (2001) 1139-1142.

Galtier, N., Piganeau, G., Mouchiroud, D. and Duret, L.: GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. Genetics 159 (2001) 907-911.

Giannelli, F., Anagnostopoulos, T. and Green, P.M.: Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B. American Journal of Human Genetics 65 (1999) 1580-7.

Gillespie, J.H.: The Causes of Molecular Evolution. Oxford University Press, Oxford, 1991.

Goldman, N. and Yang, Z.H.: Codon-based model of nucleotide substitution for protein-coding dna sequences. Molecular Biology and Evolution 11 (1994) 725-736.

Graur, D. and Li, W.H.: Fundamentals of molecular evolution, 2nd Edition ed. Sinauer Associates, Inc., 2000.

Gu, X. and Li, W.H.: A model for the correlation of mutation-rate with GC content and the origin of GC-rich isochores. Journal of Molecular Evolution 38 (1994) 468-475.

Haldane, J.B.S.: The rate of spontaneous mutation of a human gene. Journal of Genetics 31 (1935) 317-326.

Haldane, J.B.S.: The mutation rate of the gene for hemophilia and its segregation ratios in males and females. Ann. Eugenics 13 (1947) 262-271.

Haring, D. and Kypr, J.: No isochores in the Human chromosomes 21 and 22? Biochemical and Biophysical Research Communications 280 (2001) 567-573.

Hill, W.G. and Robertson, A.: The effect of linkage on limits to artificial selection. Genetical Research 8 (1966) 269-94.

Hirsh, A.E. and Fraser, H.B.: Protein dispensability and rate of evolution. Nature 411 (2001) 1046-1049.

Holland, P.W.H. and Garciafernandez, J.: Hox genes and chordate evolution. Developmental Biology 173 (1996) 382-395.

Holmquist, G., Gray, M., Porter, T. and Jordan, J.: Characterization of Giemsa Dark- and Light-Band DNA. Cell 31 (1982) 121-129.

Holmquist, G.P.: Chromosome bands, their chromatin flavors and their functional features. American Journal of Human Genetics 51 (1992) 17-37.

Huang, W., Chang, B.H.J., Gu, X., HewettEmmett, D. and Li, W.H.: Sex differences in mutation rate in higher primates estimated from AMG intron sequences. Journal of Molecular Evolution 44 (1997) 463-465.

Hughes, A.L. and Nei, M.: Pattern of nucleotide substitution at major histocompatibility complex class I loci: evidence for overdominant selection. Nature 335 (1988) 167-170.

Hughes, S., Zelus, D. and Mouchiroud, D.: Warm-blooded isochore structure in Nile crocodile and turtle. Molecular Biology and Evolution 16 (1999) 1521-1527.

Hurst, L.D. and Merchant, A.R.: High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proceedings of the Royal Society of London Series B-Biological Sciences 268 (2001) 493-497.

Hurst, L.D. and Smith, N.G.C.: Do essential genes evolve slowly? Current Biology 9 (1999) 747-750.

Hurst, L.D. and Williams, E.J.B.: Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. Gene 261 (2000) 107-114.

Hurst, L.D., Williams, E.J.B. and Pal, C.: Natural selection promotes the conservation of linkage of coexpressed genes. Trends in Genetics In press (2002).

International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. Nature 409 (2001) 860-921.

Jukes, T. and Cantor, C.: Evolution of protein molecules. In: Munro, H. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, 1969, pp. 21 - 123.

Kimura, M.: A simple method for estimating evolutionary rates of base subsititutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16 (1980) 111-120.

Kimura, M.: The Neutral Theory of Evolution. Cambridge University Press, Cambridge, 1983.

Kliman, R.M. and Hey, J.: Reduced natural-selection associated with low recombination in drosophila-melanogaster. Molecular Biology and Evolution 10 (1993) 1239-1258.

Kluwe, L., Mautner, V., Parry, D.M., Jacoby, L.B., Baser, M., Gusella, J., Davis, K., Stavrou, D. and MacCollin, M.: The parental origin of new mutations in neurofibromatosis 2. Neurogenetics 3 (2000) 17-24.

Koop, B.: Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. Trends in Genetics 11 (1995) 367-371.

Kuma, K., Iwabe, N. and Miyata, T.: Functional constraints against variations on molecules from the tissue-level - slowly evolving brain-specific genes demonstrated by protein-kinase and immunoglobulin supergene families. Molecular Biology and Evolution 12 (1995) 123-130.

Kumar, S. and Subramanian, S.: Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America 99 (2002) 803-808.

Lazaro, C., Gaona, A., Ainsworth, P., Tenconi, R., Vidaud, D., Kruyer, H., Ars, E., Volpini, V. and Estivill, X.: Sex-differences in mutational rate and mutational mechanism in the NF1 gene in neurofibromatosis type-1 patients. Human Genetics 98 (1996) 696-699.

Lercher, M.J. and Hurst, L.D.: Human SNP variability and mutation rate are higher in regions of high recombination. Trends in Genetics 18 (2002) 337-340.

Lercher, M.J., Williams, E.J.B. and Hurst, L.D.: Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. Molecular Biology and Evolution 18 (2001) 2032-2039.

Li, W.-H.: Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Journal of Molecular Evolution 36 (1993) 96-99.

Li, W.-H.: Molecular Evolution. Sinauer, Sundetland, Mass., 1997.

Li, W.H., Wu, C.-I. and Luo, C.-C.: Evolution of DNA sequences.. In: R.J.MacIntyre (Ed.), Molecular Evolutionary Genetics. Plenum, New York, 1985, pp. 1-94.

Lian, Z.H., Zack, M.M. and Erickson, J.D.: Parental age and the occurrence of birth defects. American Journal of Human Genetics 39 (1986) 648-660.

Magni, G.E.: Origin and nature of spontaneous mutations in meiotic organisms. Journal of cellular and comparative physiology 64 (Sup. 1) (1964) 165-172.

Magni, G.E. and von Borstel, R.: Different rates of spontaneous mutation during mitosis and meiosis in yeast. Genetics 47 (1962) 1097-1108.

Makova, D.K. and Li, W.H.: Strong male-driven evolution of DNA sequences in humans and apes. Nature 416 (2002) 624-626.

Marais, G., Mouchiroud, D. and Duret, L.: Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 5688-5692.

Matassi, G., Sharp, P. and Gautier, C.: Chromosomal location and evolution rate in mammalian genes, EMBO Symposium, Genomes and Chromosomes, Heidelberg, Germany, 1994.

Matassi, G., Sharp, P.M. and Gautier, C.: Chromosomal location effects on gene sequence evolution in mammals. Current Biology 9 (1999) 786-791.

McCormick, P.J., Danhauser, L.L., Rustum, Y.M. and Bertram, J.S.: Changes in ribo- and deoxyribonucleoside triphophate pools within the cell cycle of

a synchronized mouse fibroblast cell line. Biochimica et Biophysica Acta 755 (1983) 36-40.

McVean, G.T. and Hurst, L.D.: Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature 386 (1997) 388-392.

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. and Yasunaga, T.: Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harbor Symposium of Quantitative Biology 52 (1987) 863-867.

Moloney, D.M., Slaney, S.F., Oldridge, M., Wall, S.A., Sahlin, P., Stenman, G. and Wilkie, A.O.M.: Exclusive paternal origin of new mutations in Apert syndrome. Nature Genetics 13 (1996) 48-53.

Mouchiroud, D., Gautier, C. and Bernardi, G.: Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. Journal of Molecular Evolution 40 (1995) 107-113.

Nagase, T., Nagase, M., Hirose, S. and Ohmori, K.: Mutations in fibroblast growth factor receptor 2 gene and craniosynostotic syndromes in Japanese children. Journal of Craniofacial Surgery 9 (1998) 162-170.

Nagylaki, T.: Evolution of a finite population under gene conversion. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 80 (1983) 6278-6281.

Nimmo, R. and Woollard, A.: Widespread organisation of C. elegans genes into operons: fact or fiction? Bioessays 24 (2002) 983-987.

Pal, C., Papp, B. and Hurst, L.D.: Highly expressed genes in yeast evolve slowly. Genetics 158 (2001) 927-931.

Pamilo, P. and Bianchi, N.O.: Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Molecular Biology and Evolution 10 (1993) 271-81.

Papavasiliou, F.N. and Schatz, D.G.: Cell-cycle-regulated DNA double-stranded breaks in somatic hypermutation of immunoglobulin genes. Nature 408 (2000) 216-21.

Perry, J. and Ashworth, A.: Evolutionary rate of a gene affected by chromosomal position. Current Biology 9 (1999) 987-9.

Piganeau, G., Mouchiroud, D., Duret, L. and Gautier, C.: Expected Relationship between the silent substitution rate and the GC content: Implications for the evolution of isochores. Journal of Molecular Evolution 54 (2002) 129-133.

Plas, E., Berger, P., Hermann, M. and Pfluger, H.: Effects of aging on male fertility. Experimental Gerontology 35 (2000) 543-551.

Rattray, A.J., McGill, C.B., Shafer, B.K. and Strathern, J.N.: Fidelity of mitotic double-strand-break repair in Saccharomyces cerevisiae: a role for SAE2/COM1. Genetics 158 (2001) 109-22.

Richards, F.M., Payne, S.J., Zbar, B., Affara, N.A., Fergusonsmith, M.A. and Maher, E.R.: Molecular analysis of de-novo germline mutations in the Von Hippel-Lindau disease gene. Human Molecular Genetics 4 (1995) 2139-2143.

Risch, N., Reich, E.W., Wishnick, M.M. and McCarthy, J.G.: Spontaneous mutation and parental age in humans. American Journal of Human Genetics 41 (1987) 218-248.

Roberts, P.S., Chung, J., Jozwiak, S., Dabora, S.L., Franz, D.N., Thiele, E.A. and Kwiatkowski, D.J.: SNP identification, haplotype analysis, and parental origin of mutations in TSC2. Human Genetics 111 (2002) 96-101.

Rosendaal, F.R., Brockervriends, A., Vanhouwelingen, J.C., Smit, C., Varekamp, I., Vandijck, H., Suurmeijer, T., Vandenbroucke, J.P. and Briet, E.: Sex-ratio of the mutation frequencies in Hemophilia A - estimation and metaanalysis. Human Genetics 86 (1990) 139-146.

Schiebel, K., Meder, J., Rump, A., Rosenthal, A., Winkelmann, M., Fischer, C., Bonk, T., Humeny, A. and Rappold, G.: Elevated DNA sequence diversity in the genomic region of the phosphatase PPP2R3L gene in the human pseudoautosomal region. Cytogenetics Cell Genetics 91 (2000) 224-30.

Schuffenecker, I., Ginet, N., Goldgar, D., Eng, C., Chambe, B., Boneu, A., Houdent, C., Pallo, D., Schlumberger, M., Thivolet, C. and Lenoir, G.M.: Prevalence and parental origin of de novo RET mutations in multiple endocrine neoplasia type 2A and familial medullary thyroid carcinoma. American Journal of Human Genetics 60 (1997) 233-237.

Shimmin, L.C., Chang, B.H. and Li, W.-H.: Male-driven evolution of DNA sequences. Nature 362 (1993) 745-747.

Smith, N.G., Webster, M.T. and Ellegren, H.: Deterministic mutation rate variation in the human genome. Genome Research 12 (2002) 1350-6.

Smith, N.G.C. and Hurst, L.D.: The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152 (1999a) 661-673.

Smith, N.G.C. and Hurst, L.D.: The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics 153 (1999b) 1395-1402.

Spellman, P.T. and Rubin, G.M.: Evidence for large domains of similarly expressed genes in the drosophila genome. Journal of Biology 1 (2002) 5.

Strathern, J.N., Shafer, B.K. and McGill, C.B.: DNA synthesis errors associated with double-strand-break repair. Genetics 140 (1995) 965-72.

Tadokoro, R., Fujita, M., Miura, H., Shirahige, K., Yoshikawa, H., Tsurimoto, T. and Obuse, C.: Scheduled conversion of replication complex architecture at replication origins of Saccharomyces cerevisiae during the cell cycle. Journal of Biological Chemistry 277 (2002) 15881-9.

Theirry, J.-P., Macaya, G. and Bernardi, G.: An analysis of Eukaryotic Genomes by Density Gradient Centrifugation. Journal of Molecular Biology 108 (1976) 219-235.

Tiemann-Boege, I., Navidi, W., Grewal, R., Cohn, D., Eskenazi, B., Wyrobek, A.J. and Arnheim, N.: The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect. Proceedings of the National Acadamy of Sciences of the United States of America (2002) 232568699.

Williams, E.J.B. and Hurst, L.D.: The proteins of linked genes evolve at similar rates. Nature 407 (2000) 900-903.

Williams, E.J.B. and Hurst, L.D.: Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. Journal of Molecular Evolution 54 (2002a) 511-518.

Williams, E.J.B. and Hurst, L.D.: Is the synonymous substitution rate in mammals gene-specific? Molecular Biology and Evolution 19 (2002b) 1395-1398.

Williams, E.J.B., Pal, C. and Hurst, L.D.: The molecular evolution of signal peptides. Gene 253 (2000) 313-322.

Wilson, A.C., Carlson, S.S. and White, T.J.: Biochemical evolution. Annual Review of Biochemistry 46 (1977) 573-639.

Wolfe, K.: Mammalian DNA Replication: Mutation biases and the Mutation rate. Journal of Theoretical Biology 149 (1991) 441-451.

Wolfe, K.H. and Sharp, P.M.: Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. Journal of Molecular Evolution 37 (1993) 441-456.

Wolfe, K.H., Sharp, P.M. and Li, W.-H.: Mutation rates differ among regions of the mammalian genome. Nature 337 (1989) 283-285.

Yang, Z.H. and Nielsen, R.: Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution 17 (2000) 32-43.

Yeadon, P.J. and Catcheside, D.E.: Polymorphism around cog extends into adjacent structural genes. Current Genetics 35 (1999) 631-7.

Yin, L., Seri, M., Barone, V., Tocco, T., Scaranari, M. and Romeo, G.: Prevalence and parental origin of de novo RET mutations in Hirschsprung's disease. European Journal Of Human Genetics 4 (1996) 356-358.

Zorio, D.A.R., Cheng, N.S.N., Blumenthal, T. and Spieth, J.: Operons as a Common Form of Chromosomal Organization in C- Elegans. Nature 372 (1994) 270-272.

Chapter 2: The proteins of linked genes evolve at similar rates

Elizabeth J.B. Williams and Laurence D. Hurst (2000)

Nature, 407: **900-902**

coveralls, boots, hats and latex gloves. Drill bits, media, washes and syringes were sterilized by autoclaving. All autoclaving was carried out for 40 min at 121 °C in a double door pass through autoclave, continually serviced under annual maintenance contracts with quarterly inspection and testing. Drill bits, forceps and syringes used for samplings were individually packaged before sterilization. All media were placed into test tubes (no more than 8 ml per tube) and autoclaved. The manufacturer's specifications indicate that all of the autoclaved materials were sterilized to a sterility assurance level equal to 1×10^{-9}, or 1 chance of contamination in 10^9. In addition, all media to be used for samples were placed into an incubator at 37 °C for 3 days before use. If any medium batch contained even a single contaminated tube, the entire batch was discarded.

Crystals were sterilized as described[24]. All sterilants and brine washes were placed into sterile covered beakers. The cleaned crystals were immersed in 10 M NaOH for 5 min, washed in sterile saturated salt brine for 2 min, then immersed in 10 M HCl for 5 min. Following the HCl, each crystal was washed in sterile saturated salt brine buffered with sodium carbonate. This protected the stainless steel of the laminar flow hoods and drill bits from the HCl. All washes and sterilants were changed after each crystal. Sterilized crystals were removed from the buffered brine inside a dedicated Class IIA laminar flow hood fitted with an HEPA filter. The laminar flow hood was disinfected with germicidal ultraviolet light for 2 h before use. All hood surfaces and non-autoclavable equipment were disinfected with a commercial disinfectant between each drilling[24]. Sterile crystals were tightly held in a pinch clamp and penetrated with a sterile 0.5 mm wire drill operated with a micromanipulator. The inclusion samples were extracted from the crystals using sterile 25- or 250-μl syringes. The brine samples were inoculated into the sterilized growth medium.

To be sure that the spore-forming isolate was not a contaminant, the various materials used in the handling, drilling and extraction of fluids from salt crystals were deliberately contaminated with a 36-h culture of isolate 2-9-3 containing a mixture of cells and spores. All contaminated materials were streaked onto plates of tryptic-soy agar (TSA), and CAS agar amended with either 8% (CAS-8) or 20% (CAS-20) NaCl. One plate of each medium was used. All materials were then packaged as normal for crystal work, autoclaved and, after cooling, were again streaked onto TSA, CAS-8 and CAS-20 plates. All plates were incubated at 37 °C and scored for the presence of growth after 3 and 7 days. The clamp used to hold crystals during drilling in the biosafety hood, was disinfected with Wescodyne and streaked onto plates of TSA, CAS-8 and CAS-20 medium. Two plates each of TSA, CAS-8 and CAS-20 were opened and placed into the running biosafety hood following 1 h of ultraviolet light. The plates were exposed for at least 50 min, equal to the time needed to sample three crystals. The stock culture of 2-9-3 used for contamination was also streaked onto TSA, CAS-8 and CAS-20 medium before and after autoclaving.

Received 15 November 1999; accepted 4 July 2000.

1. Grant, W. D., Gemmell, R. T. & McGenity, T. J. Halobacteria: the evidence for longevity. *Extremophiles* **2**, 279–287 (1998).
2. Cano, R. J. & Borucki, M. Revival and identification of bacterial spores in 25 to 40 million year old Dominican amber. *Science* **268**, 1060–1064 (1995).
3. Denner, E. B. M. *et al.H alococcus salifodinae* sp. nov , an archaeal isolate from an Austrian salt mine. *Int. J. Syst. Bacteriol.* **44**, 774–780 (1994).
4. Huval, J. H. & Vreeland, R. H. in *General and Applied Aspects of Halophilic Bacteria.* Vol. 201 (ed. Rodriguez-Valera, F.) 53–62 (Plenum, New York, 1991).
5. Norton, C. F., McGenity, T. J. & Grant, W. D. Archaeal halophiles (halobacteria) from two British salt mines. *J. Gen. Microbiol.* **139**, 1077–1081 (1993).
6. Greenblatt, C. L. *et al.* Diversity of microorganisms isolated from amber. *Microb. Ecol.* **38**, 58–68 (1999).
7. Lambert, L. H. *et al. Staphylococcus succinus* sp. nov., isolated from Dominican amber. *Int. J. Syst. Bacteriol.* **48**, 511–518 (1998).
8. Vreeland, R. H. & Powers, D. W. in *Microbiolology and Biogeochemistry of Hypersaline Environments* (ed. Oren, A.) 53–74 (CRC, Boca Raton, Florida, 1999).
9. Vreeland, R. H. & Rosenzweig, W. D. in *Enigmatic and Extreme Microorganisms.* (ed. Seckbach, J.) 387–398 (Kluwer, Delft, 1998).
10. Croft, J. S. Upper Permian conodonts and other microfossils from the Pinery and Lamar Limestone Members of the Bell Canyon Formation and from the Rustler Formation, west Texas. Thesis, Ohio State Univ. (1978).
11. Walter, J. C. Paleontology of the Rustler Formation, Culberson County, Texas. *J. of Paleontol.* **27**, 679–702 (1953).
12. Renne, P. R., Steiner, M. B., Sharp, W. D., Ludwig, K. R. & Fanning, C. M. 40/39 Ar and U/Pb SHRIMP dating of latest Permian tephras in the Midland Basin Texas. *EOS* **77**, 794 (1996).
13. Renne, P. R., Sharp, W. D. & Becker, T. A. $^{40}Ar/^{39}Ar$ dating of langbeinite $[K_2Mg_2(SO_4)_3]$ in late Permian evaporites of the Salado Formation, Southeastern New Mexico, USA. *Mineral. Mag.* **62A**, 1253–1254 (1998).
14. Hardie, L. A., Lowenstein, T. K. & Spencer, R. J. The problem of distinguishing between primary and secondary features in evaporites. *Sixth Int. Symp. On Salt* **1**, 11–39 (1983).
15. Roedder, E. The fluids in salt. *Amer. Mineral.* **69**, 413–439 (1984).
16. Lowenstein, T. K. & Hardie, L. A. Criteria for recognition of salt-pan evaporites. *Sedimentol.* **32**, 627–644 (1985).
17. Lowenstein, T. K. Origin of depositional cycles in a Permian "saline giant": the Salado (McNutt zone) evaporites of New Mexico and Texas. *Geol. Soc. Am. Bull.* **100**, 592–608 (1988).
18. Holt, R. M. & Powers, D. W. Geological mapping of the air intake shaft at the Waste Isolation Pilot Plant. Report no. DOE/WIPP 90-051, 1–90 (U. S. Department of Energy, Carlsbad NM, 1990).
19. Lowenstein, T. K. & Spencer, R. J. Syndepositional origin of potash evaporites: petrographic and fluid inclusion evidence. *Am. J. Science* **290**, 1–42 (1990).
20. Holt, R. M. & Powers, D. W. in *Geological and Hydrological Studies of Evaporites in the Northern Delaware Basin for the Waste Isolation Pilot Plant (WIPP), New Mexico* (eds Powers, D. W., Holt, R. M., Beauheim, R. L. & Rempe, N.). *Geol. Soc. Am. Guidebook* **14**, 45–78 (1990).
21. Powers, D. W. & Hassinger, B. W. Synsedimentary dissolution pits of halite of the Permian Salado Formation, southeastern New Mexico. *J. Sed. Petrol.* **55**, 769–773 (1985).
22. Onstott, T. C., Mueller, C., Mikulki, K., Vicenzi, E. & Powers, D. W. $^{40}Ar/^{39}Ar$ laser microprobe dating of polyhalite from bedded, late Permian evaporites. *EOS* **76**, S285 (1995).
23. Vreeland, R. H., Piselli, A. F., McDonnough, S. & Myer, S. S. Distribution and diversity of halophilic bacteria in a subsurface salt formation. *Extremophiles* **2**, 321–331 (1998).
24. Rosenzweig, W. D., Peterson, J., Woish, J. & Vreeland, R. H. Development of a protocol to retrieve microorganisms from ancient salt crystals. *Geomicrobiol.* (in the press).
25. Vreeland, R. H., Anderson, R. & Murray, R. G. E. Cell wall and phospholipid composition and their contribution to the salt tolerance of *Halomonas elongata. J. Bacteriol.* **160**, 879–883 (1984).
26. Arahal, D. R, Marquez, M. C., Volcani, B. E., Schleifer, K. H. & Ventosa, A. *Bacillus marismortui* sp. nov., a new moderately halophilic species from the Dead Sea. *Int. J. Syst. Bacteriol.* **49**, 521–530 (1999).
27. Heyndrickx, M. *et al. Virgibacillus*: a new genus to accommodate *Bacillus pantothenticus* (Proom and Knight 1950). Emended description of *Virgibacillus pantothenticus. Int. J. Syst. Bacteriol.* **48**, 99–106 (1998).
28. Garabito, M. J., Arahal, D. R., Mellado, E., Marquez, M. C. & Ventosa, A. *Bacillus salexigens* sp. nov., a new moderately halophilic *Bacillus* species. *Int. J. Syst. Bacteriol.* **47**, 735–741 (1997).
29. Waino, M., Tindall, B. J., Schumann, P. & Ingvorsen, K. *Gracilibacillus* gen nov., with description of *Gracilibacillus halotolerans* gen. nov. sp. nov.; transfer of *Bacillus dipsosauri* to *Gracilibacillus dipsosauri* comb. nov., and *Bacillus salexigens* to the genus *Salibacillus* gen. nov., as *Salibacillus salexigens* comb. nov. *Int. J. Syst. Bacteriol.* **49**, 821–831 (1999).
30. Cournoyer, B. & Lavire, C. Analysis of *Frankia* evolutionary radiation using *glnII* sequences. *FEMS Microbiol. Lett.* **177**, 29–34 (1999).

Acknowledgements

The authors acknowledge the following people who helped obtain the crystal samples for this research: D. Belski, N. Rempe, R. Carrasco, T. Garcia, D. Acevedo, S. Britain, E. Keyser, B. Kinsall, A. Morin and T. Padilla. This research was supported by the US National Science Foundation: Life in Extreme Environments Program (EAR Lexen).

Correspondence and requests for materials should be addressed to R.H.V. (e-mail: rvreeland@wcupa.edu).

The proteins of linked genes evolve at similar rates

Elizabeth J. B. Williams & Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK

Much more variation in the rate of protein evolution occurs than is expected by chance[1]. But why some proteins evolve rapidly but others slowly is poorly resolved. It was proposed, for example, that essential genes might evolve slower than dispensable ones[2], but this is not the case[3]; and despite earlier claims[4], rates of evolution do not correlate with amino-acid composition[5]. A few patterns have been found: proteins involved in antagonistic co-evolution (for example, immune genes[3,6], parasite antigens[7] and reproductive conflict genes[8–10]) tend to be rapidly evolving, and there is a correlation between the rate of protein evolution and the mutation rate of the gene[1,11,12]. Here we report a new highly statistically significant predictor of a protein's rate of evolution, and show that linked genes have similar rates of protein evolution. There is also a weaker similarity of rates of silent site evolution (see ref. 13), which appears to be, in part, a consequence of the similarity in rates of protein evolution. The similarity in rates of protein evolution is not a consequence of underlying mutational patterns. A pronounced negative correlation between the rate of protein evolution and a covariant of the recombination rate indicates that rates of protein evolution possibly reflect, in part, the local strength of stabilizing selection.

To examine the effects of linkage on rates of evolution, we established a data set of rates of evolution at both non-synonymous (K_a) and synonymous (K_s) sites of mouse–rat orthologues with confirmed genomic location. Rates of evolution of genes were compared with those less than 1 centiMorgan (cM) apart, this being equivalent to on average about 2,000 kilobases (kb) in the mouse genome[14]. We calculated the modular difference (ΔK) between the K values (that is, either K_a or K_s, or K_a/K_s depending

49

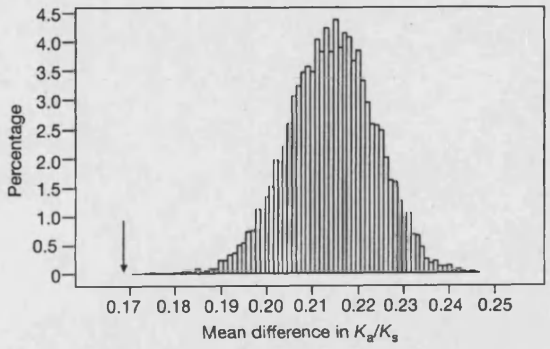


Figure 1 The distribution of 10,000 mean modular differences for K_a/K_s data. Each of the 10,000 values is the mean for a randomized version of the real K_a/K_s values, allowing in each 176 pairs of values. The actual mean modular difference for loci less than 1 cM apart is shown by an arrow.

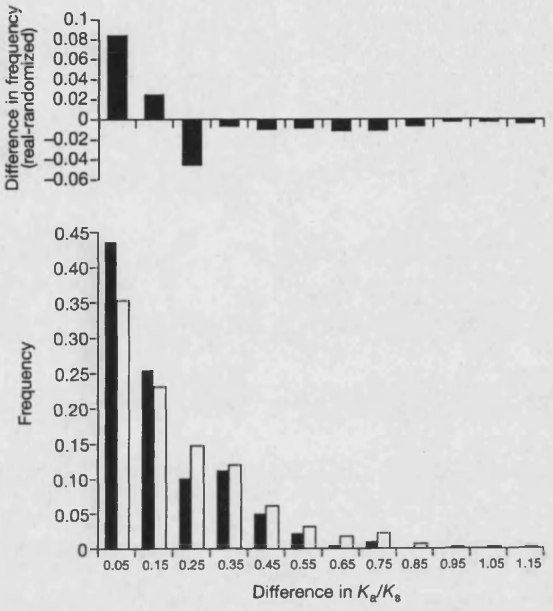


Figure 2 The frequency of differences in K_a/K_s between linked genes. Lower panel indicates the frequency of differences in K_a/K_s in 176 pairwise comparisons between linked genes (black bars), and in randomized data (white bars). Upper panel indicates the difference between the frequencies in the real and the randomized data. Numbers on the *x* axis refer to the median value of that bin (that is, 0.15 refers to values between 0.1 and 0.2).

on the analysis) for each of 176 pairs of linked genes (from 266 genes), and then determined the average.

To analyse the significance of this mean value, we followed a randomization method. We produced for each analysis 10,000 randomized data sets. In each random data set, we took all the K values from the actual data set and re-allocated them as pairs at random. For each randomized set, we calculated a mean modular difference in rate between the pairs. The proportion of the 10,000 random data sets with lower mean modular difference is a direct estimate of the P value that can be attached to the hypothesis that linked genes have similar rates of evolution.

With respect to the absolute rate of protein evolution (K_a) in 10,000 randomized data sets, none has a lower mean modular difference than that found in the real data set ($P < 0.0001$). There is also a local similarity of synonymous rates of evolution: only 163 of 10,000 random data sets have lower mean modular difference ($P = 0.016$). If the synonymous rate reflects the mutation rate, as often believed, this suggests that there may be differences between genomic regions in the underlying mutation rate (see also ref. 15). The data set is purged of tandem duplicates so neither concerted evolution nor genic non-independence can explain these patterns.

As a gene's mutation rate is likely to determine, in part, its rate of protein evolution (which explains, in part, the K_a by K_s correlation[12]), it is possible that the correlation in the rate of protein evolution of linked genes is a result of underlying mutational patterns alone. Comparing modular differences in the rate of protein evolution controlling for underlying mutation rate differences (that is, K_a/K_s), however, we find that of 10,000 random data sets, none shows a mean modular difference lower than (or even close to) that of the real data ($P < 0.0001$) (Fig. 1). The similarity of K_a/K_s of linked genes can also be demonstrated by comparing the distribution of the real 176 $\Delta K_a/K_s$ values with that from the randomized data (Fig. 2). This demonstrates an excess of small differences and a dearth of large differences in the real data (Mann–Whitney U-test: $P < 0.0001$).

By contrast, the synonymous rate bands can be explained in part as a mutational consequence of the protein rate bands. A non-synonymous mutation may be associated with a silent mutation at the adjacent site, for mechanistic reasons[16]. The spread of the non-synonymous mutation (by drift or selection) then takes the silent mutation at the adjacent site with it. Such tandem substitutions explain much of the covariation of K_a and K_s (ref. 12). To examine whether tandem substitutions might explain the synonymous rate domains, we removed them from all aligned sequences and re-calculated K_a and K_s. Although their removal reduces the mean number of synonymous sites per gene by only 3%, we find that the synonymous rate domains disappear ($P = 0.08$), while the protein rate domains remain very evident ($P < 0.0001$).

Tandem substitutions need not be the only factor contributing to the pattern. As methylation density varies around the genome[17], and as methylated CpG dinucleotides mutate about 10–20 times faster than other sites[18], we also thought that the K_s similarity may be a result of differences in methyl-induced mutation rates. This may be the case, but the effect is not profound: after removal of CpG→TpG mutations at silent sites, the local correlation of silent rates of genes less than 1 cM apart is weaker but not greatly so ($P = 0.05$ rather than $P = 0.016$). Only 2% of silent sites were removed in this analysis.

The murid genome therefore appears to be split into protein rate domains, in which the proteins of linked genes have rates of evolution that are much more similar than expected by chance, even controlling for mutation rate differences. We estimate these rate domains to be only very few cM long; we base this assertion on the fact that when we compare genes that are between 1 and 2 cM apart the genes are no longer significantly similar in terms of K_a/K_s ($P = 0.10$). This value is based on 92 pairwise comparisons between 153 genes. Were the rate bands of the order of ~10 cM, then comparing genes 1–2 cM apart would still mostly be comparing genes within a band and so should have revealed P values of the order of those obtained in the less than 1-cM class.

What might explain the existence of protein rate domains? We propose two hypotheses. First, genes of comparable function tend to have comparable rates of evolution; for example, immune genes tend to be fast evolving, neurone-specific genes tend to be slow evolving[3]. If functionally similar genes tend to cluster, then this might explain the pattern. We shall leave a test of this hypothesis to future work.

Second, we thought that these domains might reflect differences

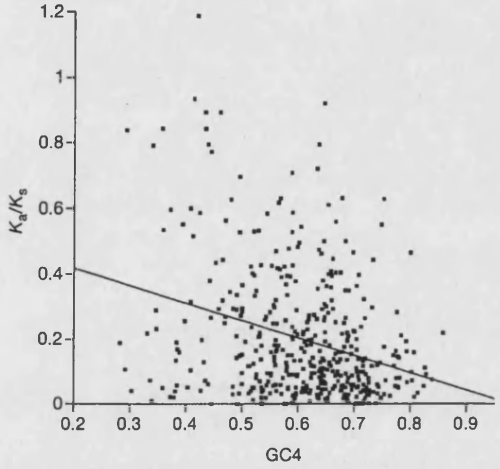


Figure 3 The covariance of the rate of protein evolution, controlling for the underlying mutation rate, and the GC4 content in the exons of the genes: K_a/K_s= 0.53–0.54 GC4, $R^2 = 0.09$, $P < 0.0001$.

in the strength of stabilizing selection. The local strength of stabilizing selection covaries with the local recombination rate[19], which varies considerably around the mouse genome[20]. Conveniently, the G + C content at fourfold redundant sites (GC4) is also thought to correlate with the recombination rate in mice[21]. This we can confirm. If recombination and G + C content are related, then we expect the Y chromosome to have a low G + C content (it never recombines), the X chromosome to have a higher one (it recombines in females only) and the autosomes to have the highest percentage of G + C. The Y chromosome has a mean G + C content at degenerate sites of 34% (n = 7), the X of 51% (n = 37) and the autosomes 61% (n = 433). These values are all highly significantly different from each other (Mann–Whitney U-test: Y < X, P = 0.0005; X < A, $P < 0.00001$). The values for the X and Y chromosomes cannot be explained as a correlate to hemizygous expression as imprinted genes (n = 15) have a G + C composition comparable to that of autosomal genes at 64%.

If, then, the local similarity of rates is due to variation in the intensity of stabilizing selection around the genome, both K_a and K_a/K_s should negatively correlate with GC4. We find a pronounced negative correlation between GC4 and both K_a and K_a/K_s for autosomal genes ($P < 0.0001$, $R^2 = 0.09$, $n = 431$) (Fig. 3). However, there may also be a difference in the sorts of genes in G + C-rich and G + C-poor isochores[22], which might in part explain the pattern.

Our results contrast with those from a previous study that examined the local similarity of rates of evolution of genes. Comparing human with rodent sequences[13], it was reported that the K_s of linked genes was highly similar, but that the similarity of K_a was of marginal significance. When we apply our methods to 176 randomly selected pairs in this rodent–human comparison, we find the same qualitative pattern, that is, strong local similarity in K_s ($P = 0.0013$) but not in K_a ($P = 0.02$). The discrepancy therefore appears to be neither due to differences in randomization methodology nor sample size. The greater distance in the rodent–human comparison as compared with the more recent mouse–rat split might possibly underlie this discrepancy.

The protein rate bands are not the only sort of mosaicism seen in the murid genome. The genome is, for example, split into G/R bands—regions that appear to correspond to late and early replicating DNA[23]. Furthermore, the mammalian genome is a mosaic of isochores, blocks of DNA within which the proportion of the bases G and C at non-coding sites (introns, third positions in codons, intergene spacer) is fairly uniform[22,24]. However, murids do not have such a well-defined isochore structure[25]. Nonetheless, considering the genes 1 cM apart, only 16 in 10,000 random data sets have lower mean modular difference in GC4 than the real linked genes ($P = 0.0016$), indicating that isochores do still exist (see also ref. 13).

Isochores and protein rate bands may be different things. Whereas the protein rates bands are not evident in genes between 1 and 2 cM ($P = 0.1$ compared with $P < 0.0001$ at less than 1 cM), the GC4 similarity of the same genes at this distance remains almost as profound as between 0 and 1 cM ($P = 0.0037$ rather than $P = 0.0016$). This also weakly indicates that isochores may be longer than usually considered: rough estimates[22] put the upper limit at about 1,300 kb, and we are still finding strong similarity at 3,000 kb (assuming the mean distance in the 1–2 cM range is 1.5 cM). There is considerable variation in the relationship between the genetic and physical maps, however, so absolute conclusions cannot be reached at present. It remains, therefore, to be more fully resolved whether protein rate bands, isochores and G bands are at all inter-related. □

Methods

We compiled a data set of mouse and rat orthologues from scrutiny of entries in HOVERGEN[26]. Genes were accepted as orthologues if, and only if, the mouse rat sequences had no other non-rodent sequence between them and at least one non-rodent sequence appeared as a sister group. This resulted in a data set of in excess of 500 gene pairs.

Each of the mouse genes was then inspected at LocusLink (www.ncbi.nlm.nih.gov/LocusLink/), by using its accession number, to establish mouse chromosomal location. These chromosomal locations are the same as those described at Mouse Genome Informatics. Those without location specified to the centiMorgan or on the X chromosome were eliminated. (X-linked genes have unusually low synonymous rates of evolution[27]). This resulted in a data set of 456 genes. Pairwise Blast searching was used to eliminate tandem duplicates from the data set. Any similarity between linked genes led to the elimination of one of the two. This resulted in a data set of 433 autosomal genes; of these, 266 had at least one neighbour within 1 cM.

We used GENETRANS to automatically extract complete coding sequences. DNA alignments were carried out by PILEUP using the default settings. The alignments were checked by eye and modified where necessary. Two genes could not be aligned adequately and were excluded. We estimated substitution rates using a described method[28] with modifications[29], and applying Kimura's two-parameter method to correct for multiple hits. For each orthologous mouse–rat gene, we therefore obtained values for the rate, per site, for both non-synonymous (K_a) and synonymous (K_s) substitutions. We also calculated the rate of protein evolution controlling for the underlying mutation rate (K_a/K_s).

In the comparison of linked genes, for most loci there was only one neighbour within 1 cM. When there was more than one, the data from any given gene was never used more than twice. In the randomized sets, if a K value was used twice in the pairwise tests, that value was also used twice in the random set. This conserves the number of pairwise comparisons and is also conservative as it permits, in the random set, a difference of zero. The programs for calculating GC4, removal of CpG→TpG mutations, and removal of doublet mutations are in Tcl script.

Received 22 February; accepted 2 June 2000.

1. Wolfe, K. H. & Sharp, P. M. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37**, 441–456 (1993).
2. Wilson, A. C., Carlson, S. S. & White, T. J. Biochemical evolution. *Ann. Rev. Biochem.* **46**, 573–639 (1977).
3. Hurst, L. D. & Smith, N. G. C. Do essential genes evolve slowly? *Curr. Biol.* **9**, 747–750 (1999).
4. Graur, D. Amino-acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.* **22**, 53–62 (1985).
5. Tourasse, N. J. & Li, W. -H. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* **17**, 656–664 (2000).
6. Kuma, K., Iwabe, N. & Miyata, T. Functional constraints against variations on molecules from the tissue-level—slowly evolving brain-specific genes demonstrated by protein-kinase and immunoglobulin supergene families. *Mol. Biol. Evol.* **12**, 123–130 (1995).
7. Hughes, A. L. Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. *Genetics* **127**, 345–353 (1991).
8. Tsaur, S. C. & Wu, C. I. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. *Mol. Biol. Evol.* **14**, 544–549 (1997).
9. Hurst, L. D. & McVean, G. T. Do we understand the evolution of genomic imprinting? *Curr. Opin. Genet. Dev.* **8**, 701–708 (1998).
10. Wyckoff, G. J., Wang, W. & Wu, C. I. Rapid evolution of male reproductive genes in the descent of man. *Nature* **403**, 304–309 (2000).
11. Makalowski, W. & Boguski, M. S. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J. Mol. Evol.* **47**, 119–121 (1998).
12. Smith, N. G. C. & Hurst, L. D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153**, 1395–1402 (1999).
13. Matassi, G., Sharp, P. M. & Gautier, C. Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786–791 (1999).

51

14. Silver, L. M. *Mouse Genetics* (Oxford Univ. Press, Oxford, 1995).
15. Casane, D., Boissinot, S., Chang, B. H. J., Shimmin, L. C. & Li, W. H. Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45**, 216–226 (1997).
16. Averof, M., Rokas, A., Wolfe, K. & Sharp, P. M. Evidence for a high frequency of double-nucleotide substitutions. *Science* **287**, 1283–1286 (2000).
17. Caccio, S. *et al.* Methylation patterns in the isochores of vertebrate genomes. *Gene* **205**, 119–124 (1997).
18. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
19. Nordborg, M., Charlesworth, B. & Charlesworth, D. The effect of recombination on background selection. *Genet. Res.* **67**, 159–174 (1996).
20. Nachman, M. W. & Churchill, G. A. Heterogeneity in rates of recombination across the mouse genome. *Genetics* **142**, 537–548 (1996).
21. Eyre-Walker, A. Recombination and mammalian genome evolution. *Proc. R. Soc. London Ser. B* **252**, 237–243 (1993).
22. Bernardi, G. The human genome: organization and evolutionary history. *Annu. Rev. Genet.* **29**, 445–476 (1995).
23. Drouin, R., Holmquist, J. P. & Richer, C. L. High-resolution replication bands compared with morphologic G- and R-bands. *Adv. Hum. Genet.* **22**, 47–115 (1994).
24. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958 (1985).
25. Galtier, N. & Mouchiroud, D. Isochore evolution in mammals: A human-like ancestral structure. *Genetics* **150**, 1577–1584 (1998).
26. Duret, L., Mouchiroud, D. & Gouy, M. HOVERGEN—a database of homologous vertebrate genes. *Nucleic Acid Res.* **22**, 2360–2365 (1994).
27. Smith, N. G. C. & Hurst, L. D. The causes of synonymous rate variation in the rodent genome: Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics* **152**, 661–673 (1999).
28. Li, W. -H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**, 96–99 (1993).
29. Pamilo, P. & Bianchi, N. O. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.* **10**, 271–281 (1993).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank W.-H. Li for comments on an earlier version of the manuscript. L.D.H. is funded by The Royal Society.

Correspondence and requests for materials should be addressed to L.D.H. (e-mail: l.d.hurst@bath.ac.uk).

Metapopulation dynamics of bubonic plague

M. J. Keeling* & C. A. Gilligan†

* *Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK*
† *Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK*

Bubonic plague is widely regarded as a disease of mainly historical importance; however, with increasing reports of incidence[1–3] and the discovery of antibiotic-resistant strains of the plague bacterium *Yersinia pestis*[4], it is re-emerging as a significant health concern[5,6]. Here we bypass the conventional human-disease models, and propose that bubonic plague is driven by the dynamics of the disease in the rat population. Using a stochastic, spatial metapopulation model, we show that bubonic plague can persist in relatively small rodent populations from which occasional human epidemics arise, without the need for external imports. This explains why historically the plague persisted despite long disease-free periods, and how the disease re-occurred in cities with tight quarantine control. In a contemporary setting, we show that human vaccination cannot eradicate the plague, and that culling of rats may prevent or exacerbate human epidemics, depending on the timing of the cull. The existence of plague reservoirs in wild rodent populations has important public-health implications for the transmission to urban rats and the subsequent risk of human outbreaks.

Large-scale human epidemics of bubonic plague have been recorded throughout history, from Roman times to the pandemic in the early 1900s. This disease has had a major social and demographic effect[7–10]; its arrival in Europe in 1348 led to the death of around one-third of the human population, and even today bubonic plague kills people in many areas of the world[1–3]. Historical data, from a variety of locations, show occasional large outbreaks of plague separated by long disease-free periods, and yet the disease clearly persists[7–9]. Understanding persistence is a common problem in general epidemic modelling[11–13], and for bubonic plague it is a central historical question. Previous models of bubonic plague have been highly anthropocentric, modelling the disease as if it were transmitted solely within human populations[10,14,15]. But consideration of the biology shows that bubonic plague is primarily a disease of rodents that is spread by fleas and only occasionally infects humans; such a disease is termed a zoonosis. From this perspective, we formulate an epizootic (animal-based disease) model for the rat and flea populations, and by coupling this with a standard epidemic (human disease) model, we identify epidemic patterns and the circumstances in which the disease causes a large number of human cases.

The life cycle of the plague can be partitioned into four stages. (1) Fleas feeding on an infected rat ingest the bacteria causing bubonic plague, and soon become infectious. (2) When an infected rat dies, its fleas leave to search for a new host. (3) The fleas usually find other rats, infect them, and so spread the disease through the rodent community. (4) Only when the density of rats is low are the fleas

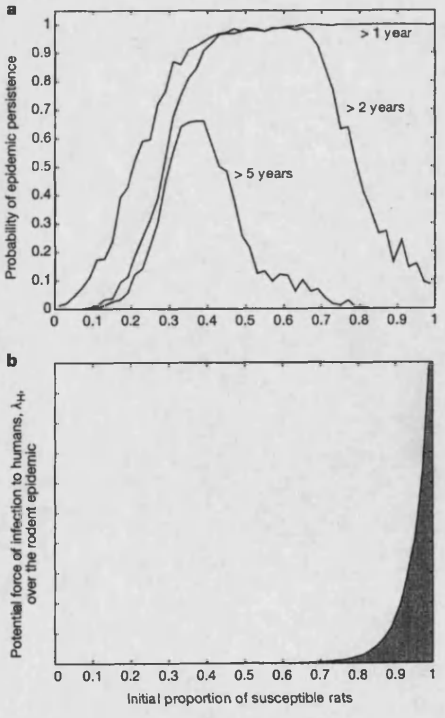


Figure 1 Results from 250 simulations of the stochastic epizootic model of bubonic plague with a rat population of 2,500. **a**, The probability that an infectious import generates an epidemic/endemic that lasts for more than 1, 2 or 5 years in the rat population. If the disease persists for more than 5 years it is likely to be in the endemic state. **b**, The potential force of infection for humans over the entire outbreak in rats ($\int_0^\infty \lambda_H(t)dt$), measured as the total number of infectious fleas that fail to find a suitable rodent host and may therefore bite and infect humans.

Chapter 3: Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: Implications for understanding the mechanistic bias of the male mutation bias

Martin J. Lercher, Elizabeth J. B. Williams and Laurence D. Hurst (2001)

Molecular Biology and Evolution, 18(11): **2032 – 2039**

Local Similarity in Evolutionary Rates Extends over Whole Chromosomes in Human-Rodent and Mouse-Rat Comparisons: Implications for Understanding the Mechanistic Basis of the Male Mutation Bias

Martin J. Lercher, Elizabeth J. B. Williams, and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, England

The sex chromosomes and autosomes spend different times in the germ line of the two sexes. If cell division is mutagenic and if the sexes differ in number of cell divisions, then we expect that sequences on the X and Y chromosomes and autosomes should mutate at different rates. Tests of this hypothesis for several mammalian species have led to conflicting results. At the same time, recent evidence suggests that the chromosomal location of genes on autosomes affects their rate of evolution at synonymous sites. This suggests a mutagenic source different from germ cell replication. To correctly interpret the previous estimates of male mutation bias, it is crucial to understand the degree and range of this local similarity. With a carefully chosen randomization protocol, local similarity in synonymous rates of evolution can be detected in human-rodent and mouse-rat comparisons. However, the synonymous-site similarity in the mouse-rat comparison remains weak. Simulations suggest that this difference between the mouse-human and the mouse-rat comparisons is not artifactual and that there is therefore a difference between humans and rodents in the local patterns of mutation or selection on synonymous sites (conversely, we show that the previously reported absence of a local similarity in nonsynonymous rates of evolution in the human-rodent comparison was a methodological artifact). We show that linkage effects have a long-range component: not one in a million random genomes shows such levels of autosomal heterogeneity. The heterogeneity is so great that more autosomes than expected by chance have rates of synonymous evolution comparable with that of the X chromosome. As autosomal heterogeneity cannot be owing to different times spent in the germ line, this demonstrates that the dominant determiner of synonymous rates of evolution is not, as has been conjectured, the time spent in the male germ line.

Introduction

If cell division is mutagenic and if the number of germ cell divisions is larger in males than in females (as is likely in most mammals), then we expect males to be the dominant source of point mutations. As Miyata et al. (1987) noted, if synonymous mutations are neutral, we can estimate the extent of the male bias to the sex ratio of mutation rates (α) by comparing the rates of synonymous evolution on the X and Y chromosomes and autosomes, as they spend different times in the germ lines of the sexes. Just this method has been applied for flies (Bauer and Aquadro 1997), rodents (Chang et al. 1994; Chang and Li 1995; Li et al. 1996; McVean and Hurst 1997; Smith and Hurst 1999*a*), cats (Slattery and O'Brien 1998), birds (Ellegren and Fridolfsson 1997), and primates (Shimmin, Chang, and Li 1993, 1994; Chang, Hewett-Emmett, and Li 1996; Li et al. 1996; Huang et al. 1997; Nachman and Crowell 2000).

It is regularly claimed (Chang et al. 1994; Chang, Hewett-Emmett, and Li 1996; Crow 1997, 2000) that the figures so obtained for the extent of the male bias correspond to the differences in the number of germ cell divisions. These claims are, however, controversial (Hurst and Ellegren 1998), because (1) we are uncertain of what the expected ratio of germ cell divisions is in most lineages, not least because estimates are highly sensitive to assumptions about the age of male reproduction (Hurst and Ellegren 1998); (2) some estimates of α from primates fall out of the range of even the lowest estimates (Bohossian, Skaletsky, and Page 2000); and (3) in rodents, the figure appears to be dependent on the sequence comparison (while X-Y comparison [Chang et al. 1994; Chang and Li 1995; Li et al. 1996] gives $\alpha = 2$, comparison of X with autosomes [McVean and Hurst 1997; Smith and Hurst 1999*a*] suggests $\alpha \gg 2$, and possibly even infinity). It seems important, then, to find other methods to determine whether we can be confident that figures for α derived using Miyata's method provide unbiased estimates of the male mutation bias and the ratio of germ cell divisions.

The critical assumption of Miyata et al. (1987) is that any difference in evolutionary rate between the X chromosome, the Y chromosome, and autosomes is attributable to different times spent in the germ lines of both sexes. However, it is also reported that along autosomes, there are regional differences in the rates of synonymous evolution (Casane et al. 1997; Matassi, Sharp, and Gautier 1999; Williams and Hurst 2000). These within-autosome effects cannot result from differences in the times spent in the male germ line. If regional effects were associated with a considerable heterogeneity of autosomal rates, this would then cast serious doubt on the validity of the method. We therefore ask about the size of the domain of local similarity. Such information should also prove helpful in resolving the causes of the regionality of rates of evolution. Furthermore, if autosomal heterogeneity is great compared with the difference between the X chromosome and autosomes, we can be confident that there is a potent force other than germ cell replication affecting synonymous substitution rates. We estimate the extent of this effect

Key words: evolutionary rate, linkage, chromosomal heterogeneity, male mutation bias.

Address for correspondence and reprints: Martin Lercher, Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom. E-mail: m.j.lercher@bath.ac.uk.

Mol. Biol. Evol. 18(11):2032–2039. 2001

54

below. To the same end, we examine how the X chromosome's rate of synonymous evolution compares with the rates of the slowest-evolving autosomes. If the X chromosome is not an outlier, we cannot be confident that the figures for α dominantly reflect the relative numbers of germ cell divisions in the two sexes.

However, prior to establishing these patterns, it is important to clarify the method, not least to understand the basis of discrepancies between previous analyses. In a recent paper on orthologous genes in the mouse and the rat, Williams and Hurst (2000) reported that linked genes show significantly similar nonsynonymous rates of evolution (K_a). While this local similarity was unexpected for nonsynonymous rates, it had long been argued that mutation rate might vary along chromosomes (Sueoka 1962; Filipski 1987; Casane et al. 1997; Nachman and Crowell 2000). Assuming that synonymous sites are not under selective pressure in mammals (Wolfe, Sharp, and Li 1989; Eyre-Walker 1991; see, however, Eyre-Walker 1999), this can be tested by searching for local similarity in the synonymous rate of evolution (K_s). However, only marginally significant similarity in K_s was found by Williams and Hurst (2000) in the mouse-rat comparison. In contrast to these results, Matassi, Sharp, and Gautier (1999) found strong local similarity of synonymous rates of evolution for human-rodent orthologs but failed to detect significant similarity of nonsynonymous rates. Thus, at present, the nature and extent of local similarities in rates of evolution are unclear and appear to be heavily dependent on the species comparison employed.

However, these discrepancies may simply reflect methodological artifacts, rather than biologically important differences. Most notably, the statistical protocols used in the recent literature on linkage effects may not be optimally suited to the detection of local similarities. Matassi, Sharp, and Gautier (1999) used a test function that summed over all pairs of genes situated within 1 cM of each other, allowing some genes to contribute multiple times. As the number of gene pairs in a linked cluster increases quadratically with cluster size, this protocol gives more weight to genes within larger clusters. A few large clusters of genes can thus dominate the test function, reducing the effective sample size and obscuring weak local similarities. Williams and Hurst (2000) circumvented this problem by pairing each gene with at most only two near neighbors. However, this pairing may be arbitrary, and part of the available information is disregarded. In the present study, we employ a method that avoids both problems. Each gene that does have linked neighbors contributes to the test function once. Its evolutionary rate is compared with the mean rate of all neighbors, thereby using all available information.

Materials and Methods

The Human-Rodent Data Set

Orthologous human and murid gene pairs were taken from the data set compiled by Duret and Mouchiroud (2000), which is accessible at http://pbil.univ-lyon1.fr/datasets/Duret_Mouchiroud_1999/data.html. Coding sequences were aligned using CLUSTAL W (Thompson, Higgins, and Gibson 1994). The numbers of substitutions per site at synonymous sites (K_s) and nonsynonymous sites (K_a) were computed with Li's (1993) method, with correction for multiple hits according to Kimura's (1980) two-parameter model. We also estimated evolutionary distances with the maximum-likelihood method introduced by Goldman and Yang (1994), implemented in the PAML package (Yang 1997). We refer to the results obtained with Li's (1993) protocol as K_a and K_s, while the distances calculated with the maximum-likelihood method are denoted by K_a^{ML} and K_s^{ML}. We found mean evolutionary distances ($\pm$SD) of $K_a = 0.073 \pm 0.071$ and $K_s = 0.50 \pm 0.14$. As evolutionary distances K_a and K_s are proportional to the corresponding evolutionary rates, we use the terms "evolutionary rate" and "evolutionary distance" interchangeably.

The gene positions on the mouse genetic map were retrieved from LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/). Duplicate genes on mouse chromosomes were identified by BLAST analysis with the default parameters for pairwise BLAST at NCBI (http://www.ncbi.nlm.nih.gov/gorf/bl2.html, score $\geq$ 39). This was done under the assumption that genes duplicated after the divergence of rodents and primates would show significant sequence similarity, detected by BLAST. We eliminated all but one copy of multicopy genes on the same chromosome. This resulted in a final data set of 1,311 autosomal and 67 X-linked human-rodent orthologs with known positions on the genetic mouse map.

We also obtained physical positions on the October 7, 2000, build of the human genome (http://genome.ucsc.edu). (On average, 1.3 cM on human chromosomes corresponds to 1 Mb; Yu et al. 2001.) This resulted in a data set of 1,849 autosomal and 80 X-linked human-rodent orthologs with known positions on the physical human map.

Expression profiles of the genes in the data set were obtained by matching expressed sequence tag (EST) data to the coding sequences (Duret and Mouchiroud 2000). For the analysis excluding immune-specific genes, we used only genes with known expression in at least one nonimmune tissue. This reduced the sample sizes to 929 orthologs on mouse autosomes and 1,545 orthologs on human autosomes.

The Mouse-Rat Data Set

A data set of mouse-rat orthologs was collected by scrutinizing entries in Hovergen (the Homologous Vertebrate Gene Database, available at http://www.hgmp.mrc.ac.uk; Duret et al. 1994). Genes were considered orthologs if the gene family tree contained no internal nonrodent branch between the mouse and rat sequence branches and if at least one nonrodent sequence appeared as an outgroup to the mouse and rat sequences. This resulted in a data set of over 500 gene pairs.

Each mouse gene was inspected at LocusLink (www.ncbi.nlm.nih.gov/LocusLink/) via its accession number to establish mouse chromosomal location. These chromosomal locations were identical to those described

55

at Mouse Genome Informatics (www.informatics.jax.org; Mouse Genome Informatics—The Jackson Laboratory 1996). Those without locations specified to the centimorgan and those on the X chromosome were eliminated from the data set. Pairwise BLAST searching was used to eliminate tandem duplicates from the data set. This resulted in a data set of 475 autosomal genes.

GENETRANS (GCG program suite at HGMP, http://hgmp.mrc.ac.uk) was used to automatically extract complete coding sequences. DNA alignments were carried out with PILEUP (also part of GCG) using the default setting. The alignments were checked by eye and modified if necessary. Substitutions per site were estimated as described for the human-rodent data set. We found mean evolutionary distances (±SD) of K_a = 0.036 ± 0.038 and K_s = 0.17 ± 0.05.

Statistics

For each gene, we calculated the difference between K_a (K_s) and the mean of all its neighbors within a certain distance range. The mean absolute difference was calculated by summing over all genes. We then created a set of 100,000 random mean differences by permuting the K_a (K_s) values of all genes at random. To test for within-chromosome local similarity, we permuted only genes within the same chromosome. To test if local similarity was caused by a covariation of the rates with local GC content, we swapped only genes within classes of similar GC contents at third codon positions (GC3). Each GC3 class contained 10% of the full data set. We defined a measure of the local similarity in K_s (and analogous in K_a) as the ratio of two mean absolute differences in K_s, i.e., of the observed (linked) difference and the difference expected without linkage effects (from randomization):

$$\rho_s = \frac{\text{observed } K_s \text{ difference}}{\text{expected } K_s \text{ difference}}.$$

Thus, a value of $\rho_s = 0.85$ means that on average the difference between the synonymous rates of linked genes is only 85% of the difference expected by chance.

To compare our results with those obtained by Matassi, Sharp, and Gautier (1999), we defined a second test function as the mean squared K_a (K_s) difference of all linked pairs. This measure is equivalent to the I statistics employed by Matassi, Sharp, and Gautier (1999). The contribution of each gene to this test function is proportional to the number of its neighbors. A corresponding random distribution was created as above.

Chromosomal heterogeneity was measured with the test function

$$\chi^2 = \sum_{i=1}^{i=22} \frac{(K_i - K_{\text{mean}})^2}{v_i},$$

where K_i is the mean K_s for chromosome number i and v_i is the expected variance of K_s on the same chromosome, derived from $v_i = v/N_i$ (v is the variance of K_s in the full data set, and N_i is the number of genes on the chromosome). We can test the hypothesis that there is heterogeneity by creating 1,000,000 randomized data sets and asking how many have χ^2 values greater than that seen in the real data set. In each randomization run, genes were randomly reassigned to chromosomes, keeping only the total number of genes per chromosome intact. The distribution created by this randomization procedure is approximately χ^2-distributed (data not shown).

Results

Linked Genes Have Similar K_a Values

Linked genes were defined as those within 1 cM of each other on the genetic mouse map. Employing the same statistical protocol as Matassi, Sharp, and Gautier (1999), we confirmed that with this method no significant local K_a similarity is found for human-rodent orthologs (P = 0.90). However, the protocol amplifies the influence of localized clusters of genes, which increases the variance of the randomized data sets. Weighting genes more evenly by using the mean K_a difference as defined in *Materials and Methods*, we found highly significant K_a similarity for linked genes (ρ_a = 0.917, P = 0.00010). This result was robust to the removal of immune-specific genes (P = 0.0026). In the mouse-rat comparison, linked genes also had significantly similar K_a values when analyzed with this protocol (ρ_a = 0.860, P = 0.0023). When linked genes on human autosomes were defined as those within 1 Mb of each other, we also found significant local similarity in K_a values (ρ_a = 0.946, P = 0.0027).

Linked Genes Have Similar K_s Values

In comparing the mean K_s difference of the human-rodent data set with randomized genomes, we found highly significant similarity for linked genes on mouse autosomes (ρ_s = 0.859, $P < 10^{-5}$). The same significance was obtained after removal of immune-specific genes. On human autosomes, we found highly significant similarity for linked genes within 1 Mb of each other (ρ_s = 0.857, $P < 10^{-5}$). In the mouse-rat comparison, K_s similarity was not significant (ρ_s = 0.941, P = 0.087). However, when we included more gene comparisons by extending the definition of linked genes to those within 5 cM of each other, local similarity reached significance (ρ_s = 0.935, P = 0.024). Nonetheless, there seems to be a discrepancy in the strengths of the local similarity between the two species comparisons.

Rate Similarities Extend over Whole Chromosomes

Within this wider linkage definition of d = 5 cM, K_s similarity in the human-rodent comparison was still highly significant. What is the range of this local similarity? For any examined linkage radius (d = 1, 2, 5, 20, and 200 cM), we found highly significant "local" K_s similarity ($P < 10^{-5}$). A range of d = 200 cM includes all genes residing on the same mouse chromosome. Thus, the K_s similarity of linked genes extends over all genetic length scales, from 1 cM up to whole chromosomes. In comparing mean K_s values on human autosomes, we also found very extensive heterogeneity between autosomes. As seen in table 1 and figures 1 and

Table 1
Chromosomal Heterogeneity of Mouse Autosomes in Evolutionary Rates, Calculated According to Li's (1993) Method (K_a, K_s) and with Maximum-Likelihood (K_a^{ML}, K_s^{ML})

	K_a	K_s	K_a^{ML}	K_s^{ML}
Human-rodent (human autosomes)	$\chi^2 = 32.8$, $P = 0.048$	$\chi^2 = 148.4$, $P < 10^{-6}$	$\chi^2 = 32.6$, $P = 0.051$	$\chi^2 = 227.7$, $P < 10^{-6}$
Human rodent (mouse autosomes)	$\chi^2 = 33.6$, $P = 0.014$	$\chi^2 = 76.6$, $P < 10^{-6}$	$\chi^2 = 32.1$, $P = 0.021$	$\chi^2 = 86.5$, $P < 10^{-6}$
Mouse-rat (mouse autosomes)	$\chi^2 = 32.9$, $P = 0.024$	$\chi^2 = 31.7$, $P = 0.023$	$\chi^2 = 36.5$, $P < 0.01$	$\chi^2 = 27.8$, $P > 0.05$

2, not one in a million random human or mouse genomes has more chromosomal K_s heterogeneity than the real data. The local K_s similarity in the mouse-rat comparison was much weaker and was not detectable on all length scales. However, when we tested for chromosomal heterogeneity, we found again that mouse chromosomes had significantly different mean K_s values (table 1).

Does such heterogeneity also exist for nonsynonymous rates of evolution? In both species comparisons, we found significant heterogeneity of mean autosomal K_a, which for the human-rodent comparison was evident both on the mouse map and on the human map (table 1). Thus, genes positioned on the same autosome have significantly similar rates of nonsynonymous and synonymous site evolution in both the human-mouse and the mouse-rat comparisons. In the more distant human-rodent comparison, we found the chromosomal effect for K_a to be much weaker than that for K_s.

Because of its importance for understanding previous conflicting results on male bias to the mutation rate, we further analyzed K_s heterogeneity among human autosomes. We found that it was robust to analysis of only those genes with known expression in nonimmune tissue ($\chi^2 = 113.9$, $P < 10^{-6}$). It was also robust to analysis of K_s after codons involved in doublet substitutions were removed to reduce the covariance of synonymous and nonsynonymous rates (Smith and Hurst 1999*b*; Duret and Mouchiroud 2000) ($\chi^2 = 142.8$, $P < 10^{-6}$), showing that the effect is not owing to heterogeneity in rates of nonsynonymous evolution (such heterogeneity was only marginally significant; table 1). Eight of the human autosomes (numbers 4, 13, 14, 15, 16, 17, 19, and 21) showed significant ($P < 0.05$ from 1,000,000 random permutations) deviation of mean K_s from null expectations (fig. 1). Four (numbers 4, 14, 17, and 19) remained significant after Bonferroni correction. Similarly, seven mouse autosomes (numbers 2, 4, 5, 8, 10, 11, and 12) showed significant deviation from null expectations (fig. 2); four (numbers 4, 8, 10, and 11) remained significant after Bonferroni correction.

That genes have not resided on the same autosome for all of the evolution between rodents and humans makes our analysis conservative: rearrangements should act as a randomizing process, tending to homogenize rates of evolution between autosomes. This constitutes evidence against germ cell mutations as the dominant determiner of substitution rates: the germ cell division model for the male mutation bias explicitly fails to predict between-autosome heterogeneity, as all autosomes spend the same time in the male germ line.

Autosomal Heterogeneity Is So Great That the Human X Chromosome Is Not an Outlier

If time spent in the male germ line were the dominant predictor of K_s, then the X chromosome should appear as an outlier. However, the human X chromosome is not an outlier. While the X chromosome has the lowest mean K_s, we find that two autosomes (with a total of 184 genes) have K_s values almost as low (fig. 1). In simulations, not one out of a million randomly rearranged genomes had at least 184 genes on autosomes with error bars overlapping those of the X chromosome ($P < 10^{-6}$). Thus, while time spent in the male germ

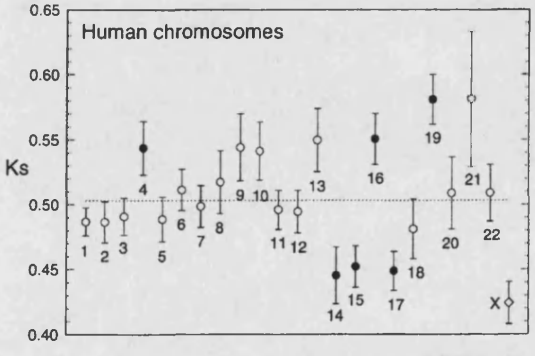


FIG. 1.—Mean rate of evolution at synonymous sites (K_s) for the 22 human autosomes and the X chromosome. Eight autosomes (shown as black dots) show significantly high or low rates of evolution under a null model in which all autosomes evolve on average at the same rate ($P < 0.05$ from randomization data). The dotted line shows mean K_s on autosomes (0.490 ± 0.003).

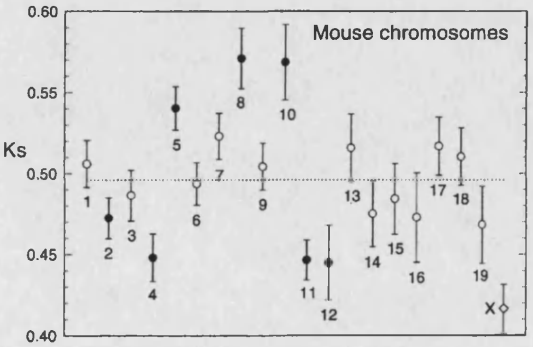


FIG. 2.—Mean rate of evolution at synonymous sites (K_s) from the human-rodent comparison for the 19 mouse autosomes and the X chromosome. Seven autosomes (shown as black dots) show significantly high or low rates of evolution. The dotted line shows mean K_s on autosomes (0.496 ± 0.004).

57

Table 2
Ratios of Observed and Randomized Rate Differences (ρ) and Corresponding *P* Values for Within-Chromosome Local Similarity (i.e., controlling for between-chromosome effects)

	Human-Rodent				Mouse-Rat			
	K_a		K_s		K_a		K_s	
Distance (cM)	ρ_a	P	ρ_s	P	ρ_a	P	ρ_s	P
0–2	0.936	0.00042	0.925	$<10^{-5}$	0.866	0.00089	0.952	0.15
2–4	0.968	0.19	0.944	0.011	1.006	0.72	0.969	0.29
4–6	0.992	0.22	0.956	0.019	—	—	—	—
6–8	0.995	0.28	0.989	0.47	—	—	—	—

line may well contribute to the variance in K_s, it fails to explain its majority.

Within-Chromosome Local Similarity Exists Independent of Chromosomal Effects

In our randomization protocol, we can control for between-chromosome heterogeneity by swapping rate values only between genes on the same chromosome. We find that in addition to chromosomal effects, there is local (within-chromosome) K_s similarity, although this is significant only in the human-rodent comparison and not in the mouse-rat comparison. Table 2 shows the results from measuring K_a and K_s similarities for human-rodent orthologs in ever-expanding rings (i.e., among genes between 0 and 2 cM apart, between 2 and 4 cM apart, etc.). Up to around 6 cM on the mouse map, local within-chromosome K_s similarity persists in the human-rodent comparison. As can be seen from table 2, there is also local within-chromosome similarity in K_a. However, this similarity is short-ranged, and no significant within-chromosome K_a similarity could be detected for genes farther than 2 cM apart in both species comparisons. On the human map, we could not detect significant local similarity in K_a or K_s beyond a distance of 2 Mb (distance 0–2 Mb: $\rho_a = 0.949$, $P = 0.00068$; $\rho_s = 0.896$, $P < 10^{-5}$; distance 2–4 Mb: $\rho_a = 0.984$, $P = 0.23$; $\rho_s = 0.987$, $P = 0.25$).

GC Content Does Not Explain Local Similarity

The genomes of vertebrates have been described as mosaics of long (>300 kb) DNA segments homogeneous in base composition, termed isochores (Bernardi 1995). While the existence of distinct isochores has recently been questioned, there are strong local similarities in GC content (International Human Genome Sequencing Consortium 2001). It is also known that evolutionary rates are influenced by GC content, although the exact form of this dependence is still a matter of debate (Wolfe, Sharp, and Li 1989; Bernardi, Mouchiroud, and Gautier 1993; Smith and Hurst 1999*b*; Bielawski, Dunn, and Yang 2000; Hurst and Williams 2000). One can then hypothesize that it is similarity in GC content—and not linkage as such—which leads to local rate similarities. This hypothesis can be tested by a randomization protocol that permutes genes only within classes of similar GC3 (Matassi, Sharp, and Gautier 1999). We still find highly significant similarity in K_s ($\rho_s = 0.860$, $P < 10^{-5}$; and $\rho_s = 0.861$, $P < 10^{-5}$) and in K_a ($\rho_a = 0.917$, $P = 0.00021$; and $\rho_a = 0.953$, $P = 0.0092$) for human-rodent orthologs within 1 cM on the mouse map and within 1 Mb on the human map, respectively. Thus, the local similarity in evolutionary rates is not a consequence of similarity in GC content.

Maximum-Likelihood Estimates Confirm the Local Rate Similarities

The protocol used for the estimation of evolutionary rates might influence our ability to pick up the similarities discussed above. To test this hypothesis, we repeated our calculations using rates obtained with the maximum-likelihood protocol introduced by Goldman and Yang (1994). In accordance with the above results, we found highly significant similarity for genes within 1 cM on the mouse map, both in K_a^{ML} (human-rodent: $\rho_a = 0.914$, $P = 0.00004$; mouse-rat: $\rho_a = 0.850$, $P = 0.0032$) and in K_s^{ML} (human-rodent: $\rho_s = 0.776$, $P < 10^{-5}$). As before, the similarity in K_s^{ML} for the mouse-rat comparison was not significant within the chosen range of 1 cM ($\rho_s = 0.936$, $P = 0.085$). Again, both similarities extended over whole chromosomes, leading to chromosomal heterogeneity in K_a^{ML} and in K_s^{ML} (see table 1). However, the heterogeneity in K_a^{ML} was now just below significance for human autosomes, and heterogeneity in K_s^{ML} was not significant for the mouse-rat comparison. The range of the within-chromosome similarity on the mouse map was unchanged compared with that reflected in table 2. Due to the allowance for biased composition, the maximum-likelihood estimate of K_s^{ML} depends much more on GC3 (Smith and Hurst 1999*b*) compared with the K_s value obtained with Li's (1993) protocol. However, when permuting human-rodent orthologs within classes of similar GC3, we still found significant local similarity on the mouse map (K_s^{ML}: $\rho_s = 0.802$, $P < 10^{-5}$; K_a^{ML}: $\rho_a = 0.915$, $P = 0.00030$) and on the human map (K_s^{ML}: $\rho_s = 0.87$, $P < 10^{-5}$; K_a^{ML}: $\rho_a = 0.95$, $P = 0.0053$).

Low K_s Similarity in Rodents Is Not Due to Small Sample Size or Evolutionary Distance

The relative strengths of the K_a and K_s similarities were very different in both species comparisons. Whereas regionality in K_s was very strong in the human-rodent comparison (see also Matassi, Sharp, and Gautier 1999), it was weak in the mouse-rat comparison (see also Wil-

liams and Hurst 2000). For local K_a similarity, the situation was reversed: it was strong in the mouse-rat comparison but could be detected only by carefully weighting gene clusters in the human-rodent comparison; indeed, it was not reported by Matassi, Sharp, and Gautier (1999). The latter discrepancy we have shown to be an artifact of methodology. However, the weak K_s similarity in the mouse-rat comparisons is not so obviously artifactual. Repeatedly drawing 475 random genes from the human-rodent data set, we found higher local K_s similarity (within 1 cM) than in the mouse-rat comparison ($P < 0.09$) in 978 out of 1,000 draws. This is significant evidence that sample size does not fully explain the different strengths in the two species comparisons ($P = 0.022$). What, then, are the causes of this difference?

It has been conjectured that selection is less efficient in rodents (Bernardi 1995), e.g., because of small effective population sizes in structured subpopulations or increased mutation rates. If spatial structure in K_s is maintained by selection (e.g., on codon usage, GC content, or modifiers of the mutation rate), a reduction in local similarity in rodents would then be expected. The reduced heterogeneity of local GC content in murids compared with other mammals has been cited as evidence for such a reduction in selective pressure (Bernardi 2000).

An alternative hypothesis for the discrepancy in local K_s similarities asserts no difference in the evolutionary mechanisms acting in humans and in rodents, but assumes the difference in divergence time to be the underlying cause. Two species that diverged as recently as the mouse and the rat have accumulated few mutations. The variances in K_s and K_a are thus small, but the variance in estimates of K_s are dependent on gene size and may be proportionally large. For K_s, this sampling variance may drown any linkage effects that would otherwise be visible. However, due to varying selective pressures, nonsynonymous rates of evolution have much higher underlying variances (as a percentage of the mean; see the figures given in *Materials and Methods*) than synonymous rates. Here, the sampling variance can be considered relatively small and does not obscure local effects. Following this line of argument, we expect the relative standard deviation (as a percentage of the mean) to be higher in the mouse-rat comparison. However, we find the same relative standard deviations for our K_s estimates in both species comparisons (human-rodent, 28%; mouse-rat, 29%; this is unchanged when immune-specific genes are excluded).

We performed a set of Monte Carlo simulations to distinguish between the two alternative explanations of the different strengths in local K_s similarity. We characterized each human-rodent ortholog by its number of synonymous sites (as determined with the maximum-likelihood method) and by its mutation rate, which we approximated by K_s^{ML}. All sequences were "evolved" repeatedly with a Poisson process (i.e., under a strictly neutral model, with independence of substitutions, and with no substitutional bias). We found that the combined effects of Poisson noise (due to short evolutionary distance) and small sample size appear insufficient to explain the observed discrepancy in local K_s similarity between the two species comparisons.

Discussion

Genes within a few centimorgans of each other have similar rates of synonymous and nonsynonymous evolution. This similarity is found in comparisons of closely related (mouse-rat), as well as more distant (human-rodent), mammalian species. To detect the similarity optimally in randomization protocols, one has to carefully consider the statistical treatment of linked clusters.

We confirmed that the local similarity in synonymous rate (K_s) was much weaker in the mouse-rat comparison than in the human-rodent comparison. The smaller sample size and the shorter evolutionary time over which mutational processes acted in the mouse-rat comparison should both add relative noise. However, our simulations show that the combined effect is highly unlikely to obscure the local K_s similarity to the extent seen in our data. We must thus conclude that there exist real underlying differences in the spatial patterns of mutation or selection on synonymous sites between humans and rodents. Corresponding differences are known to exist in compositional genome organization and have been attributed to weakened selection in the rodent genome (Bernardi 2000).

Local similarities of evolutionary rates are detectable on two different genetic length scales: within a few centimorgans, and on whole chromosomes. What light do these results shed on previous estimates of the male mutation bias from comparisons of rates on the X and Y chromosomes and autosomes? Here, we found (1) unexpectedly extensive variance between autosomes in rate of synonymous gene evolution and (2) that the X-linked genes have a rate of evolution comparable to that of genes on some autosomes. Neither finding can be explained as an artifact of sampling, as our randomization tests are robust to such problems. As location on the X chromosome is highly conserved among species (all X-linked genes in our study were X-linked in both humans and rodents), a translocation of sequences from the X chromosome to autosomes also cannot be responsible. These results are not consistent with the hypothesis that time spent in the male germ line is the dominant determiner of synonymous rates of evolution. If we may suppose that the synonymous rate is a measure of the mutation rate, as is consistent with the equality of synonymous rates and intronic rates of evolution (Smith and Hurst 1998) and patterns of codon usage bias (Eyre-Walker 1991), then this suggests that germ cell division is not the dominant cause of mutation. In summary, comparison of rates of synonymous evolution on the X and Y chromosomes and autosomes cannot be assumed to be an unbiased method for determining male mutation bias and the relative proportion of germ cell divisions in the two germ lines.

These results may also be helpful in interpreting some of the previous discrepant estimates of the male

59

mutation bias. Notably, if we suppose there to be some other mutagenic force (e.g., recombination) whose effects differ within and between autosomes and also between the X chromosome, the Y chromosome, and autosomes, then we should expect that X-Y comparisons and the X-autosome comparisons need not provide the same estimate for α. Such a lack of concordance has been observed in rodents (McVean and Hurst 1997; Smith and Hurst 1999*a*). Furthermore, the extent of the regional heterogeneity is so great that sampling from one region alone is likely to lead to biased estimates. The recently obtained unusually low estimate ($\alpha = 1.7$) found for primates (Bohossian, Skaletsky, and Page 2000) came from analysis of only one block of sequence. The discrepancy between this and prior estimates was conjectured to reflect a difference between coding and noncoding regions, but this now appears unlikely, as a sample of pseudogenes (Nachman and Crowell 2000) provides a higher estimate ($\alpha \approx 4$). The most likely cause of this discrepancy, we wish to suggest, is a biased estimate owing to genomic regionality in rates of evolution, as we have described here.

What causes the local similarities described in this paper? We can reject similarities in GC content, which have been put forward as a possible explanation, as a likely cause. Given a high rate of synonymous evolution of genes in the pseudoautosomal region (Perry and Ashworth 1999), we might predict that one component of the variation might be mutations induced by recombination (the pseudoautosomal region being a region with an unusually high recombination rate). This would be in line with evidence from yeast (Strathern et al. 1995) and from mammalian somatic hypermutation (Papavasiliou and Schatz 2000), suggesting that repair of double-strand breaks, possibly during recombination, is mutagenic. The hypothesis does not obviously concur, however, with the finding that in fruit flies, in which males do not undergo recombination, point mutations appear to be as commonly derived from males as from females (Bauer and Aquadro 1997). Further analysis of this issue in mammals will require construction of adequate recombinational maps in which ancestral recombination rates can be estimated. This we leave to future work. It will also be very valuable to know which types of point mutations are typically induced by faulty repair of double-strand breaks.

The range of the local similarity in nonsynonymous rates of evolution appears to be much smaller than that for synonymous rates. Still, we found significant chromosomal heterogeneity in K_a, which cannot easily be explained in terms of biases of the repair machinery. As synonymous and nonsynonymous rates are essentially independent when calculated with the maximum-likelihood method, the local similarity in K_a does not appear to be due to the underlying mutation rate (significant chromosomal heterogeneity was also found for K_a/K_s and K_a^{ML}/K_s^{ML}; data not shown). To explain the local similarity in nonsynonymous rates, we have to invoke selective explanations. It has recently been suggested that clusters of similarly expressed genes, termed "expression modules," are responsible for the local K_a similarity (Hurst and Eyre-Walker 2000).

LITERATURE CITED

Bauer, V. L., and C. F. Aquadro. 1997. Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **14**:1252–1257.

Bernardi, G. 1995. The human genome: organization and evolutionary history. Annu. Rev. Genet. **29**:445–476.

———. 2000. Isochores and the evolutionary genomics of vertebrates. Gene **241**:3–17.

Bernardi, G., D. Mouchiroud, and C. Gautier. 1993. Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Biol. **37**:583–589.

Bielawski, J. P., K. A. Dunn, and Z. Yang. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics **156**:1299–1308.

Bohossian, H. B., H. Skaletsky, and D. C. Page. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature **406**:622–625.

Casane, D., S. Boissinot, B. H. J. Chang, L. C. Shimmin, and W. H. Li. 1997. Mutation pattern variation among regions of the primate genome. J. Mol. Evol. **45**:216–226.

Chang, B. H. J., D. Hewett-Emmett, and W.-H. Li. 1996. Male-to-female ratios of mutation-rate in higher primates estimated from intron sequences. Zool. Stud. **35**:36–48.

Chang, B. H. J., and W.-H. Li. 1995. Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked Ube-1 genes and pseudogenes. J. Mol. Evol. **40**: 70–77.

Chang, B. H. J., L. C. Shimmin, S. K. Shyue, D. Hewett-Emmett, and W.-H. Li. 1994. Weak male-driven molecular evolution in rodents. Proc. Natl. Acad. Sci. USA **91**:827–831.

Crow, J. F. 1997. The high spontaneous mutation rate: is it a health risk? Proc. Natl. Acad. Sci. USA **94**:8380–8386.

———. 2000. The origins patterns and implications of human spontaneous mutation. Nat. Rev. Genet. **1**:40–47.

Duret, L., and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. **17**:68–74.

Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN—a database of homologous vertebrate genes. Nucleic Acid Res. **22**:2360–2365.

Ellegren, H., and A. K. Fridolfsson. 1997. Male-driven evolution of DNA sequences in birds. Nat. Genet. **17**:182–184.

Eyre-Walker, A. C. 1991. An analysis of codon usage bias in mammals: selection or mutation bias? J. Mol. Evol. **33**: 442–449.

———. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics **152**:675–683.

Filipski, J. 1987. Correlation between molecular clock ticking, codon usage, fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. FEBS Lett. **271**:184–186.

Goldman, N., and Z. H. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Huang, W., B. H. J. Chang, X. Gu, D. Hewett-Emmett, and W. H. Li. 1997. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. J. Mol. Evol. **44**:463–465.

60

HURST, L. D., and H. ELLEGREN. 1998. Sex biases in the mutation rate. Trends Genet. **14**:446–452.

HURST, L. D., and A. EYRE-WALKER. 2000. Evolutionary genomics: reading the bands. BioEssays **22**:105–107.

HURST, L. D., and E. J. B. WILLIAMS. 2000. Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. Gene **261**:107–114.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

LI, W.-H., D. L. ELLSWORTH, J. KRUSHKAL, B. H. J. CHANG, and D. HEWETT-EMMETT. 1996. Rates of nucleotide substitution in primates and rodents and the generation time effect hypothesis. Mol. Phylogenet. Evol. **5**:182–187.

MCVEAN, G. T., and L. D. HURST. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature **386**:388–392.

MATASSI, G., P. M. SHARP, and C. GAUTIER. 1999. Chromosomal location effects on gene sequence evolution in mammals. Curr. Biol. **9**:786–791.

MIYATA, T., H. HAYASHIDA, K. KUMA, K. MITSUYASU, and T. YASUNAGA. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. Cold Spring Harb. Symp. Quant. Biol. **52**:863–867.

NACHMAN, M. W., and S. L. CROWELL. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**:297–304.

PAPAVASILIOU, F. N., and D. G. SCHATZ. 2000. Cell-cycle-regulated DNA double-strand breaks in somatic hypermutation of immunoglobulin genes. Nature **408**:216–221.

PERRY, J., and A. ASHWORTH. 1999. Evolutionary rate of a gene affected by chromosomal position. Curr. Biol. **9**:987–989.

SHIMMIN, L. C., B. H. CHANG, and W.-H. LI. 1993. Male-driven evolution of DNA sequences. Nature **362**:745–747.

———. 1994. Contrasting rates of nucleotide substitution in the X-linked and Y-linked zinc-finger genes. J. Mol. Evol. **39**:569–578.

SLATTERY, J. P., and S. J. O'BRIEN. 1998. Patterns of Y and X chromosome DNA sequence divergence during the Felidae radiation. Genetics **148**:1245–1255.

SMITH, N. G. C., and L. D. HURST. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. **47**:493–500.

———. 1999*a*. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics **152**:661–673.

———. 1999*b*. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics **153**:1395–1402.

STRATHERN, J. N., B. K. SHAFER, and C. B. MCGILL. 1995. DNA-synthesis errors associated with double-strand-break repair. Genetics **140**:965–972.

SUEOKA, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48**:582–592.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

WILLIAMS, E. J. B., and L. D. HURST. 2000. The proteins of linked genes evolve at similar rates. Nature **407**:900–903.

WOLFE, K. H., P. M. SHARP, and W.-H. LI. 1989. Mutation rates differ among regions of the mammalian genome. Nature **337**:283–285.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

YU, A., C. ZHAO, Y. FAN et al. (11 co-authors). 2001. Comparison of human genetic and sequence-based physical maps. Nature **409**:951–953.

ADAM EYRE-WALKER, reviewing editor

Accepted July 13, 2001

Chapter 4: Is the synonymous substitution rate in mammals gene specific?

Elizabeth J. B. Williams and Laurence D. Hurst (2002)

Molecular biology and evolution, 19(8): **1395-1398**

Letter to the Editor

Is the Synonymous Substitution Rate in Mammals Gene-Specific?

Elizabeth J. B. Williams and Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Claverton Down, United Kingdom

There is a great deal of variation in silent rates of evolution (Ks) between genes in the same species pair comparison (Bernardi, Mouchiroud, and Gautier 1993; Wolfe and Sharp 1993; Mouchiroud, Gautier, and Bernardi 1995). This may represent random fluctuation (Kumar and Subramanian 2002) or may be deterministically caused. Evidence for the latter comes from the finding that the rate of silent site evolution of a gene is repeatable across independent lineages (Bulmer, Wolfe, and Sharp 1991; Mouchiroud, Gautier, and Bernardi 1995; Bielawski, Dunn, and Yang 2000). Most notably, Mouchiroud, Gautier, and Bernardi (1995) found that the number of synonymous substitutions per synonymous site (Ks) for a gene in the human-cow comparison was a very strong predictor of the Ks of the same gene in the mouse-rat comparison.

This repeatability has been used as evidence (Mouchiroud, Gautier, and Bernardi 1995) for selection acting upon silent sites in orthologous genes. One argument holds, for example, that if purifying selection favors a particular amino acid at a given site, it should also favor the translationally most accurate codon and codon usage bias would be selected for. The repeatability of Ks and the Ka-Ks correlation are then attributed to the same selectionist cause. However, this interpretation is questionable on a number of counts. Most notably, with one possible exception (Debry and Marzluff 1994), there is no compelling evidence that codon usage bias in mammals is the result of selection (Eyre-Walker 1991; Karlin and Mrazek 1996; Urrutia and Hurst 2001). Alternative interpretations of the putative fact are also possible. These include gene- or chromosome-specific mutation rates and gene-specific rates of biased gene conversion, all of which could give repeatable Ks and a Ka-Ks correlation.

Leaving the difficulties of interpretation aside, we wish to note two potential problems with the prior analysis, that of Mouchiroud, Gautier, and Bernardi. First, these authors did not constrain the orthologs to be autosomal. Mammalian X-linked genes often have a low Ks, most probably in part because of the relatively short time spent in the male germ line (for review see Hurst and Ellegren 1998). A data set with numerous X and autosomal genes could give repeatability of Ks, but this may represent repeatability at the chromosomal level (and be mutationally driven) rather than be on account of selection on silent sites. To address this we analyze a data set of orthologs known to be autosomal.

Address for correspondence and reprints: Laurence D. Hurst, Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom.
E-mail: l.d.hurst@bath.ac.uk.

Mol. Biol. Evol. 19(8):1395–1398. 2002
 ISSN: 0737-4038

Second, it is now well established that the sophistication of estimators of the silent site rate of evolution can have a major effect on many molecular evolutionary patterns, such as the relationship between Ks and GC4 (Pesole et al. 1995; Smith and Hurst 1998; Bielawski, Dunn, and Yang 2000; Hurst and Williams 2000) and between Ka and Ks (Smith and Hurst 1998; Bielawski, Dunn, and Yang 2000). For example, even if the mutation rate does not vary as a function of GC content, many methods give an artifactual report of an inverted U-shaped distribution (Pesole et al. 1995). Mouchiroud, Gautier, and Bernardi used the LWL method, which appears to be highly prone to this bias (Pesole et al. 1995). This artifact alone could lead to apparent repeatability. Is the repeatability of the synonymous substitution rate equally sensitive to method, and might the findings of Mouchiroud, Gautier, and Bernardi (1995) be an artifact of the usage of methods that do not make good allowance for such things as biased base composition?

HOVERGEN (Release 40, May 2000; GenBank release 116) was used to collect orthologs. Human, cow, mouse, and rat orthologs were collected for each gene. A total of 150 such sets of orthologs was originally collected; however, because of poor alignments or insufficient evidence of orthology, this was reduced to a final data set of 116 ortholog sets. Using HOVERGEN is probably the best method for determining orthology of sequences. It does, however, have its shortfalls. Most notably, if a gene from one species has a faster rate of evolution than the others, then the true ortholog can fall out at the bottom of the tree because of long branch attraction. Hence, whereas the human, cow, mouse, and rat sequences might cluster as a family, the topology is not as expected. But this result could also come about were the putatively fast evolving sequence also a paralog. Hence, we must reject the families with the unusual topologies, but this may bias to finding sequences with less overdispersed rates of evolution (and hence more repeatable rates). There were, however, only six families that we had to reject because the bovine gene seemed to have a faster rate of evolution and fell out at the bottom of the phylogeny of the putative orthologs.

The coding sequences were extracted from GenBank files using GBPARSE. CLUSTALW was used to align all four translated sequences together, and the nucleotide alignments were reconstituted from the protein alignments and the nucleotide sequences. For one gene, Lamp1 (J04182, L09113, M32015, M34959), it was not possible to align the signal peptide because of a high level of degeneracy. So the mature peptide alone was aligned. Signal peptides are known to evolve faster than average; they seem to conserve hydrophobicity and little else (Williams, Pal, and Hurst 2000).

Ka and Ks were calculated using LWL (Li, Wu, and Luo 1985), Li 1993 (Li 1993), YN00 (Yang and

Table 1
The Pearson's Correlation of Human-Cow Substitution Rates with Those for Mouse-Rat Orthologs

	All Genes		>300 Codons		Autosomal		Autosomal >300 Codons		Autosomal >300 Codons No Doublets	
No.:	116		77		106		71		71	
	r^2 (%)	P	r^2 (%)	P	r^2 (%)	P	r^2 (%)	P	r^2 (%)	P
Ks LWL.....	9.5	0.001	8.7	0.009	9.1	0.002	7.7	0.019	5.5	0.049
Ks Li 1993 ..	5.0	0.016	3.0	0.130	5.1	0.020	2.6	0.180	1.1	0.394
Ks YN00....	4.4	0.032	4.0	0.090	4.3	0.045	4.1	0.102	4.9	0.069
Ks ML.....	7.9	0.002	3.0	0.130	8.1	0.003	2.9	0.156	1.7	0.274
K4 TN.....	0.2	0.660	0.1	0.795	0.3	0.558	0.1	0.777	0.0	0.957
K4 K2P.....	0.4	0.474	0.6	0.500	0.7	0.403	0.6	0.519	0.1	0.773
Ka LWL.....	62.3	0.000	70.3	0.000	63.0	0.000	69.7	0.000	69.3	0.000
Ka Li 1993 ..	71.9	0.000	78.4	0.000	71.9	0.000	77.7	0.000	74.6	0.000
Ka YN00....	69.4	0.000	77.0	0.000	69.9	0.000	76.2	0.000	72.8	0.000
Ka ML.....	71.8	0.000	77.5	0.000	72.1	0.000	76.8	0.000	72.6	0.000
GC4.....	67.5	0.000	72.4	0.000	68.8	0.000	72.9	0.000	72.9	0.000

Nielsen 2000) and ML (Goldman and Yang 1994), comparing the human gene with the cow gene and the mouse gene with the rat gene. K4 was calculated using K2P (Kimura 1980), as well as using TN (Tamura and Nei 1993). The mean GC4 content was calculated from the human and cow genes as well as the mouse and rat genes. The entire protocol from GenBank file to final result was automated using a Perl script available from the authors. The mean length of the four orthologs defined the length of the gene. The human gene location was ascertained by looking at LocusLink at the NCBI website (http://www.ncbi.nlm.nih.gov/LocusLink/index.html). Doublets were removed using a Perl script obtainable from the authors.

For all cases where we looked at a subsample of the complete data set, e.g., when removing short genes, we tested the significance of the results obtained in that subsample by creating randomly subsampled data sets of the same size as the actual subsample. In nearly all cases, the difference in significance of the subsample compared with the complete data set was not significant. Any noteworthy results will also be pointed out in the results. Otherwise it should be assumed that there is no difference.

The data set was restricted to genes definitely present on human autosomes. However, we failed to replicate the results of Mouchiroud, Gautier, and Bernardi. (see table 1). Notably, whereas they report an r^2 of 28%, our autosomal data set reports at most an r^2 value of around 8%. Possibly of importance is the finding that the highest r^2 that we found was that found employing the method used by Mouchiroud, Gautier, and Bernardi (i.e., LWL). We also found that K4, whichever method was used, gave no significant evidence of repeatability. This suggests that to some extent the previous high estimate may be a methodological artifact.

One obvious alternative explanation for the discrepancy is that our set of genes, unlike the previous one, is not constrained to longer genes in which estimates of Ks are more accurate. However, restricting analysis to genes with greater than 300 codons, we find that any repeatability all but disappears (see table 1). By contrast, Mouchiroud, Gautier, and Bernardi found the opposite tendency. The biased LWL method reports a weakly significant effect, but otherwise no method reports a significant effect. Note that for this analysis the sample size is close to that employed by Mouchiroud, Gautier, and Bernardi (their data set, $N = 85$; our long gene autosomal data set, $N = 71$).

We appear then to have failed to replicate Mouchiroud, Gautier, and Bernardi's result and can find no convincing evidence that given genes have characteristic synonymous substitution rates. One further possible cause of the discrepancy is that the prior data set took genes regardless of their chromosomal location. We then analyzed a data set of the known autosomal genes, known X-linked genes ($N = 3$), and genes of unknown location ($N = 7$). However, again, none of the results were highly significant (table 1, $N = 116$). Again, the method that gave the strongest correlation was LWL, the method implemented by Mouchiroud, Gautier, and Bernardi (1995). K4 again showed the weakest repeatability. Increasing the gene size did not increase the strength of repeatability (see table 1).

The discrepancy between the results is probably not because of a difference in the proportion of X-linked genes in the two data sets: we failed to detect a significant difference in the strength of correlation when autosomal and X-linked genes are used compared with when autosomal ones alone are used. This was established by randomly removing three genes, 100 times, from the data set and comparing the repeatability of synonymous substitutions with the data set lacking the three X-linked genes (Ks Li, $P = 0.50$; Ks LWL, $P = 0.42$; Ks YN00, $P = 0.43$; Ks ML, $P = 0.51$).

It has been hypothesized that a repeatability of Ks is expected if (1) Ka is repeatable, and (2) Ka covaries with Ks. A selectionist explanation can be provided for both findings. Is then our lack of evidence for strong repeatability possibly caused by Ka not being repeatable or to Ka and Ks not covarying? As might be expected, Ka is strongly and unambiguously gene-specific (see ta-

64

Table 2
Pearson's Correlation of Substitution Rates and GC Content in Human-Cow and Mouse-Rat Comparisons, for Genes Longer than 300 Codons

	Ka versus Ks		Ka versus Ks, Doublets Removed		Ks versus K4 TN		Ka versus K4 TN		Ks versus GC4	
Method	r^2 (%)	P	r^2 (%)	P	r^2 (%)	P	r^2 (%)	P	r^2 (%)	P
Human-Cow Comparison										
LWL	20.8	0.000	0.2	0.734	48.0	0.000	7.0	0.026	0.9	0.444
Li 1993	25.9	0.000	0.3	0.666	77.2	0.000	9.5	0.009	10.5	0.006
YN00	3.1	0.144	1.6	0.303	49.2	0.000	7.3	0.023	33.8	0.000
ML	12.8	0.002	0.5	0.575	58.7	0.000	7.7	0.019	45.2	0.000
Mouse-Rat Comparison										
LWL	5.4	0.05	0.6	0.530	56.5	0.000	1.9	0.253	5.4	0.050
Li 1993	5.5	0.049	0.1	0.799	84.1	0.000	2.0	0.235	0.7	0.502
YN00	0.2	0.711	1.6	0.295	50.8	0.000	1.5	0.316	12.2	0.003
ML	3.0	0.148	0.4	0.621	68.3	0.000	2.1	0.231	6.2	0.036

ble 1). We also find in the human-cow comparison a strong correlation between Ka and Ks in all methods except YN00 (see table 2), and a weaker correlation in the mouse-rat comparison. The correlation between Ka and K4 is much weaker, however (see table 2). Removing doublets removes the Ka-Ks correlation (see table 2) but made little difference to the evidence for repeatability of Ks (see table 1). This suggests that any repeatability of Ks found using some of the methods is not caused by whatever causes the Ka-Ks correlation.

Possibly the most notable of our findings is that the extent of repeatability is highly method-dependent. Methods that use both twofold and fourfold sites come to conclusions different from those obtained using methods that employ only fourfold sites. The latter never detect repeatability, whereas the former do under some circumstances. The estimates of K4 repeatability are not greatly affected by employing long genes alone, so sample size effects are unlikely. Why then do the different measures of Ks report such drastically different estimates for the extent of repeatability? We suggest that GC content may be of importance.

GC content is strongly repeatable between mammalian orthologs (see table 1) (Bernardi 2000). If a method is biased with respect to GC content, then the repeatability of Ks could simply be an artifact of the repeatability in GC. Importantly, Pesole et al. (1995) showed that methodology was particularly sensitive to GC content such that at extremes of GC content many methods tended to be inaccurate. Notably, they report that K4 using the TN93 correction correctly recovers GC independence where other methods (e.g., Kimura 2 parameter), such as those built into LWL, fail to account for biased base composition. We did some similar simulations and found a similar result (data not shown). We also find in our data set that with LWL, Ks shows the typical (artifactual) inverted U–shaped distribution of Ks and GC4. We can then be confident that in some part the repeatability shown using LWL is an artifact of inaccurate Ks estimation at the extremes of GC content.

The weak repeatability shown using Goldman and Yang's maximum likelihood method and its approximation (YN00) is most likely also, in part, a reflection of a relationship between GC4 and Ks. With both, as found previously (Smith and Hurst 1998; Bielawski, Dunn, and Yang 2000; Hurst and Williams 2000), we find genes with a high GC content also have a high Ks. Under both methods the number of synonymous sites rapidly decreases as third site GC content tends toward 100%. This is because twofold codons are all either GA- or TC-ending. If GC content is skewed, the method supposes that these twofold degenerate third sites largely represent sites at which only nonsynonymous substitutions can occur. Consequently, almost regardless of the number of synonymous changes, the number of synonymous changes per synonymous site must increase because the number of synonymous sites is plummeting. It is unclear whether this method is unbiased.

In order to confirm our results we performed a repeatability analysis comparing human-pig with mouse-rat orthologs. Results were much as with the analysis using cows. For example, using just autosomal genes, the repeatability in Ks calculated using Li 1993 was significant ($r^2 = 9.2\%$, $P = 0.003$, $N = 91$), but using K4 measured by Tamura and Nei, we found no evidence of repeatability ($r^2 = 1.5\%$, $P = 0.255$). As before restricting analysis to genes with greater than 300 codons, we found no evidence of repeatability using either method (Li 1993, $r^2 = 2.4\%$, $P = 0.301$, $N = 46$; K4 TN, $r^2 = 0.1\%$, $P = 0.834$). Interestingly we did find repeatability in the larger data set when we used K4 calculated using Kimura 2 parameter. This effect disappeared when we restricted the data set to longer genes. This again suggests that the difference between analyses is caused by methodological differences and GC-based artifacts.

We have been able to reach a few conclusions. We cannot under any circumstance recover the high correlation previously reported (Mouchiroud, Gautier, and Bernardi 1995) between the synonymous substitution rate of orthologs in the human-cow comparison and that in the mouse-rat comparison. In part the strength of the previous correlation may reflect methodological bias. What weak effects we can detect are only found using Ks, never found using K4 with a high parameter multihit correction method. A priori we expect K4 to be the better method because Ks has numerous potential GC-re-

lated biases. Given the uncertainty of what is fact and artifact as regards the GC-Ks correlation, it seems safest to conclude that there is no unambiguous evidence that individual autosomal mammalian genes have their own characteristic synonymous substitution rates. This is consistent with Kumar and Subramanian's (2002) finding that the variation in K4 between genes in a genome may be accounted for by a stochastic model, rather than a deterministic one.

Acknowledgments

We thank Clare Hamilton for assistance with the pig analysis and BBRSC for funding for L.D.H. We are grateful to the editor and two anonymous referees for comments on an earlier version of the manuscript.

LITERATURE CITED

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene **241**:3–17.

Bernardi, G., D. Mouchiroud, and C. Gautier. 1993. Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. **37**:583–589.

Bielawski, J. P., K. A. Dunn, and Z. H. Yang. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics **156**:1299–1308.

Bulmer, M., K. H. Wolfe, and P. M. Sharp. 1991. Synonymous nucleotide substitution rates in mammalian genes—implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA **88**:5974–5978.

Debry, R. W., and W. F. Marzluff. 1994. Selection on silent sites in the rodent H3 histone gene family. Genetics **138**: 191–202.

Eyre-Walker, A. 1991. An analysis of codon usage in mammals: selection or mutation bias? J. Mol. Evol. **33**:442–449.

Goldman, N., and Z. H. Yang. 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Hurst, L. D., and H. Ellegren. 1998. Sex biases in the mutation rate. Trends Genet **14**:446–452.

Hurst, L. D., and E. J. B. Williams. 2000. Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. Gene **261**:107–114.

Karlin, S., and J. Mrazek. 1996. What drives codon choices in human genes. J. Mol. Biol. **262**:459–472.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base subsititutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kumar, S., and S. Subramanian. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA **99**:803–808.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

Li, W.-H., C.-I. Wu, and C.-C. Luo. 1985. A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

Mouchiroud, D., C. Gautier, and G. Bernardi. 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. **40**:107–113.

Pesole, G., G. Dellisanti, G. Preparata, and C. Saccone. 1995. The importance of base composition in the correct assessment of genetic distance. J. Mol. Evol. **41**:1124–1127.

Smith, N. G. C., and L. D. Hurst. 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. **47**:493–500.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

Urrutia, A., and L. D. Hurst. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics **159**:1191–1199.

Williams, E. J. B., C. Pal, and L. D. Hurst. 2000. The molecular evolution of signal peptides. Gene **253**:313–322.

Wolfe, K. H., and P. M. Sharp. 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. **37**:441–456.

Yang, Z. H., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43.

Pekka Pamilo, reviewing editor

Accepted April 29, 2002

66

Chapter 5: Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores

Laurence D. Hurst and Elizabeth J. B. Williams (2000)

Gene, 261:**107-114**

ELSEVIER

Gene 261 (2000) 107–114

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores

Laurence D. Hurst*, Elizabeth J.B. Williams

Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

Received 19 May 2000; received in revised form 2 October 2000; accepted 13 October 2000
Received by G. Bernardi

Abstract

Many attempts to test selectionist and neutralist models employ estimates of synonymous (Ks) and non-synonymous (Ka) substitution rates of orthologous genes. For example, a stronger Ka–Ks correlation than expected under neutrality has been argued to indicate a role for selection and the absence of a Ks-GC4 correlation has been argued to be inconsistent with neutral models for isochore evolution. However, both of these results, we have shown previously, are sensitive to the method by which Ka and Ks are estimated. Using a maximum likelihood (ML) estimator (GY94) we found a positive correlation between Ks and GC4 and only a weak correlation between Ka and Ks, lower than expected under neutral expectations. This ML method is computationally slow. Recently, a new ad hoc approximation of this ML method has been provided (YN00). This is effectively an extension of Li's protocol but that also allows for codon usage bias. This method is computationally near-instantaneous and therefore potentially of great utility for analysis of large datasets. Here we ask whether this method might have such applicability. To this end we ask whether it too recovers the two unusual results. We report that when the ML and earlier ad hoc methods disagree, YN00 recovers the results described by the ML methods, i.e. a positive correlation between GC4 and Ks and only a weak correlation between Ks and Ka. If the ML method can be trusted, then YN00 can also be considered an adequately reliable method for analysis of large datasets. Assuming this to be so we also analyze further the patterns. We show, for example, that the positive correlation between GC4 and Ks is probably in part a mutational bias, there being more methyl induced CpG → TpG mutations in GC rich regions. As regards the evolution of isochores, it seems inappropriate to use the claimed lack of a correlation between GC and Ks as definitive evidence either against or for any model. If the positive correlation is real then, we argue, this is hard to reconcile with the biased gene conversion model for isochore formation as this predicts a negative correlation.

Keywords: Covariation of GC content; Silent site substitution rate; Evolution; Isochores

1. Introduction

1.1. Estimators of Ka and Ks

There are very many tests for the possible role of selection and many of these use estimates of the rates of synonymous (Ks) and non-synonymous (Ka) divergence between orthologous genes. For example, genes (or subdomains) under directional selection will typically show Ka/Ks>>1 while those under purifying selection have Ka/Ks<<1. Similarly, Ka and Ks covary, but the extent of this covariance appears to be greater than expected under neutrality thus indicating selection (for references and recent analysis see Smith and Hurst, 1999b; Bielawski et al., 2000).

The calculation of Ka and Ks can be achieved by very many different methods. Broadly these can be divided into ad hoc methods (the majority) and maximum likelihood methods (e.g. Goldman and Yang, 1994) which provide an explicit statistical defence for the protocol. The ad hoc methods differ in their assumptions with some taking account of both the transition-transversion (Ts/Tv) bias and codon usage bias (such as YN00) and others only allowing for the former (e.g. Li's protocol (Li, 1993; Pamilo and Bianchi, 1993)) or for neither. The change from the earliest methods (e.g. LWL (Li et al., 1985)) to those making better correction for Ts/Tv ratios saw a relatively dramatic change in estimates of Ks.

For a long time now these Li based protocols have been considered the norm in molecular evolutionary analyses and changes to the methodology have not resulted in qualita-

Abbreviations: Ks, synonymous; Ka, non-synonymous; ML, maximum likelihood; Ts/Tv, transition-transversion

* Corresponding author. Tel.: + 44-1225-826424; fax: +44-1225-826779.
E-mail address: l.d.hurst@bath.ac.uk (L.D. Hurst).

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00489-3

68

tively different patterns. However, we recently reported that the ML method, which also allows for codon usage bias and Ts/Tv bias, recovers at least two results that the earlier ad hoc methods, that make no allowance of codon usage bias, did not recover: a correlation between GC4 and Ks, and only a weak correlation between Ka and Ks (Smith and Hurst, 1999). The latter result is of importance in that it upsets the assertion that the Ka–Ks correlation is evidence for selection.

The former result is important in the context of the evolution of isochores. This context we here explain.

1.2. The evolution of isochores and the Ks-GC4 correlation

Most mammalian and bird genomes are believed to be divided into discrete blocks of distinct G and C content, so called isochores (Bernardi et al., 1985), that are thought to range from 300 kb to over 1000 kb in size. The existence of local similarity of GC content of linked genes in rodents and humans supports the existence of such a pattern (Matassi et al., 1999; Williams and Hurst, 2000). Understanding the evolutionary forces responsible for the evolution of isochores has been one of the foci of the debate between neutralists and selectionists, but remains unresolved.

One specific selectionist hypothesis argues that isochores are an adaptation to homeothermy, or more generally to a hot conditions. It is postulated that some regions of the genome increased GC content to ensure that DNA would be less prone to fall apart under the high temperatures seen in mammals and birds (Bernardi and Bernardi, 1986). Metaphorically, one might see GC rich bands as acting like staples holding two loosely connected fibres together at regular intervals.

The neutralist hypotheses point to some potential mutational processes that might vary around the genome such as DNA repair (Filipski, 1987), mutational bias (Suoka, 1988), changes in nucleotide pools during replication (Wolfe et al., 1989) or biased gene conversion (Eyre-Walker, 1993).

One of the most important pieces of evidence against these models is an apparent lack of correlation between GC content at four-fold degenerate sites (GC4) and the rate of silent site evolution (Ks) (Bernardi et al., 1993; Bernardi et al., 1997). For example, it was believed that the efficiency of repair might be correlated with GC content (Filipski, 1988). Early evidence that Ks is highest in regions of high AT content (Filipski, 1988; Ticher and Graur, 1989), therefore appeared to support the models. It was then demonstrated (Bernardi et al., 1993), however, that the negative correlation between GC and Ks was an artefact of low sample size and so this model was rejected.

The replication time/nucleotide pool model posits that variation in both silent substitution rate and base composition is due to systematic differences in the rate and pattern of mutation over regions of the genome, the differences arising because mutation patterns vary with the timing of replication of different chromosomal regions in the genome (Wolfe et al., 1989). It is known that different genes repeatably replicate, at least in somatic cells, at different times. Whether the timing of replication and GC content are in any way connected is more contentious with some studies showing no relationship (Eyre-Walker, 1992). However, the most recent data show that chromosomal bands containing H3 isochores (those with the highest GC content) replicate almost entirely or largely at the onset of S phase (Federico et al., 1998).

Detailed models of nucleotide misincorporation in a nucleotide gene pool changing through a replication event predict, through most parameter space, an inverted-u or inverted-v shaped relationship between GC4 and the rate of silent site evolution (Ks) (Gu and Li, 1994). Such a pattern had been found (Wolfe et al., 1989). However, it was then claimed that this too was an artefact of low sample sizes (Bernardi et al., 1993) as with a considerably larger data set no such relationship between GC4 and Ks was found. By contrast, again, in an extensive analysis of 363 mouse–rat othologs the inverted u-shaped distribution was recovered (Wolfe and Sharp, 1993). It was then counter-argued, however, that the result might be due to methodological artefacts that lead to underestimation of Ks at extremes of GC content (Mouchiroud et al., 1995).

1.3. Methodology and the Ks-GC4 correlation

It is at this point that ML analysis showing that GC4 and Ks are positively correlated (Smith and Hurst, 1999b; Bielawski et al., 2000) becomes important. The ML method should control for the artefacts that were identified. In part it is the cause of this that we wish to further explore in this paper. However, our more central concern is one of methodology. Yang and colleagues have recently developed an ad hoc approximation to the ML estimator (YN00) (Yang and Nielsen, 2000). This method cannot be considered superior to the ML method in terms of the accuracy of the estimates. However, importantly, this second method is computationally almost instantaneous whereas the ML method is very slow. YN00 is therefore potentially of great utility in the analysis of large datasets. This is only so, however, if it can be considered adequately unbiased.

In this paper therefore we ask whether the two unusual results found using the ML estimator are also found using YN00. That is, we shall ask whether we find a highly significant positive correlation between Ks and GC4 and whether the Ka–Ks correlation is weaker than typically reported. If YN00 recovers these two results, and if it can be safely assumed that the ML results are the most reliable (which has yet to be definitively established especially for the case of pair-wise contrasts), then we might reasonably recommend YN00 as the optimal method when computational time is limited.

Further, given that YN00 is conceptually similar to the previously used ad hoc methods, such as those developed by Li and colleagues (Li, 1993; Pamilo and Bianchi, 1993)

but with added considerations to take account of codon usage bias, the differences between YN00 and the earlier ad hoc methods, might reasonably be considered to be owing to biases resulting from codon usage bias, the bias that is thought to explain the depression in Ks at extremes of GC content.

Recently we also reported (Williams and Hurst, 2000) that linked genes evolve at similar rates at both non-synonymous and synonymous sites. We proposed that the similarity in protein rates might be owing to differences in the strength of stabilising selection around the genome. In support of this we showed (a) that GC content and recombination rate are positively correlated and (b) that GC and Ka (and Ka/Ks) are negatively correlated. As stabilising selection is expected to be at its most efficient in regions of high recombination (Nordborg et al., 1996), this series of correlations is consistent with the stabilising selection model. Given that YN00 produces estimates of Ks that are sensitive to GC it is worth asking whether the correlation between protein rates of evolution (Ka and Ka/Ks) and GC4 are robust. We therefore provide the first consideration of this issue using a method that makes control for codon usage bias.

2. Methods

We have assembled a dataset of 422 confirmed autosomal mouse-rat orthologs. However three we could not unambiguously align, leaving us with 419. We used autosomal genes as X-linked genes are known to have low Ks values most probably for reasons that are unrelated to the evolution of isochores (Smith and Hurst, 1999a). It is for this reason that present data set is slightly smaller than that previously analysed (Smith and Hurst, 1999b). Orthology was confirmed using HOVERGEN (Duret et al., 1994). Genes were accepted as othologs if, and only if, the mouse rat sequences had no other non-rodent sequence between them and at least one non-rodent sequence appeared as a sister group. Chromosomal location was found by searching LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/) by murine accession number. This permits ready access to gene location data at Mouse Genome Informatics.

Coding sequence was extracted automatically using the annotations in the GenBank entry. The GC content at four-fold degenerate sites within exons was calculated using a Tcl script. Alignment was done using Pileup. The method of Yang and Nielsen was applied to the alignments using the implementation in PAML2.0k (http://abacus.gene.ucl.ac.uk/software/paml.html).

The effects of CpG → TpG mutations was examined using a Tcl script that automated the removal of CG/TG pairs with C and T at the third site or CA/CG pairs with A and G at the third site.

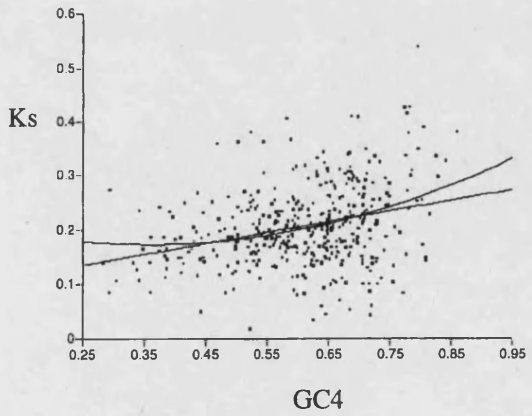


Fig. 1. The relationship between GC4 and Ks for 419 autosomal genes in the mouse-rat comparison. Ks is here calculated using the method of Yang and Nielsen (2000). The linear regression and quadratic best fit curve are both shown.

3. Results

3.1. GC4 and Ks are positively correlated

We find, as previous maximum likelihood analysis finds (Smith and Hurst, 1999b), that GC4 and Ks are signif icantly positively correlated: $Ks = 0.086 + 0.198$ GC4, $P < 0.0001$, $R^2 = 9.0\%$ (see Fig. 1). Fitting a quadratic function to the data indicates a best fit of a shallow u-shaped function with a minimum around GC4 = 0.35, not the inverted u previously claimed. There exist a few putative outliers. Removal of these does not disturb the result.

3.2. Older parametric analyses report 'typical' results

We have also applied older methods, notably that of Li. Pamilo and Bianchi (Li, 1993; Pamilo and Bianchi, 1993), and found no correlation between Ks and GC4, as Bernardi and colleagues previously reported (Bernardi et al., 1993): $Ks = 0.186 - 0.017$ GC4; $P = 0.459$, $R^2 = 0.1\%$ (Fig. 2). A quadratic fit reports a weak inverted-u shaped function (Fig. 2). These results are typical of previous analyses based on large datasets (Bernardi et al., 1993; Wolfe and Sharp, 1993; Bernardi et al., 1997; Smith and Hurst, 1999b).

Restricting analysis to the substitution rate at four-fold degenerate sites we find the same lack of correlation between Ks and GC4: $K4 = 0.182 - 0.014$ GC4, $P = 0.575$; $R^2 = 0.1\%$. The best fit quadratic is also an inverted-u shape (data not shown). Therefore we can be confident that the unusual behaviour reported in the YN00 analysis is not due to an unusual dataset.

3.3. The positive correlation between GC4 and Ks is probably due, in part, to methyl induced mutations

An obvious hypothesis to explain our findings is that we are seeing the consequences of methyl induced mutations. Methylation is probably the most potent mutagen in the

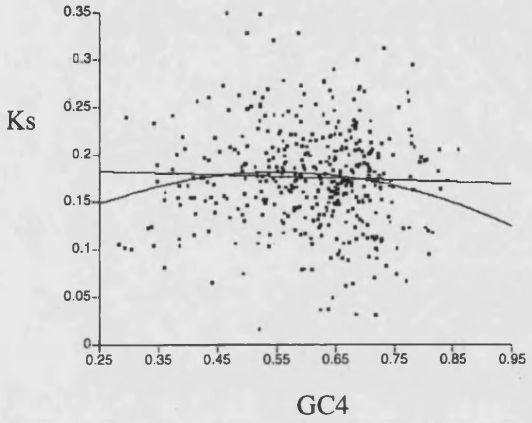


Fig. 2. The relationship between GC4 and Ks for 419 autosomal genes in the mouse-rat comparison. Ks is here calculated using the method of Li, Pamilo and Bianchi (Li, 1993; Pamilo and Bianchi, 1993). The linear regression and quadratic best fit curve are both shown.

mammalian genome and tends to convert CpG residues into TpG residues, as well as converting CpG into CpA. By chance alone, one would expect more CpG pairs in GC rich regions and hence possibly a higher Ks in GC rich regions. This hypothesis we can examine by removing CpG → TpG and CpG → CpA in the alignments, where the substitutions occur at third sites, and recalculating Ks.

The mean Ks values are reduced about 25% by removal of the methyl associated mutations, which is testament to their high frequency: using YN00 the mean Ks goes down from 0.206 ± 0.003 to 0.156 ± 0.002: using Li's protocol the mean goes from 0.176 ± 0.0026 to 0.141 ± 0.00225. The slope of the regression between Ks and GC4 using YN00 is reduced 40% from 0.198 to 0.118. Nonetheless, the slope remains significantly greater than zero: Ks = 0.084 + 0.118 GC4; $P < 0.0001$, $R^2 = 5.5\%$ (Fig. 3). The

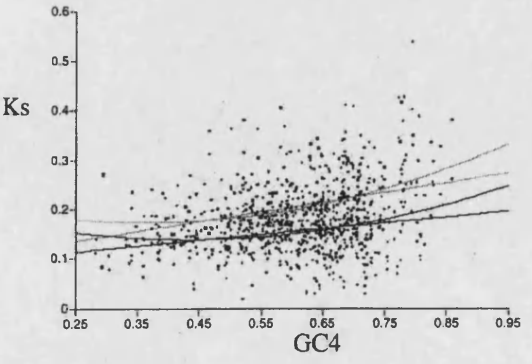


Fig. 3. The relationship between GC4 and Ks for 419 autosomal genes in the mouse-rat comparison. Ks is here calculated using the method of Yang and Nielsen (2000). The lines in grey are for the data without CpG → TpG's removed. The lines in black are for those with these methyl associated mutations removed.

best fit quadratic remains a u-shaped function. It would appear then that some, but not all of the correlation between Ks and GC4 is attributable to a mutation bias. However, this must be regarded as only a preliminary examination. The removal of some sites reduces the size of the genes being analysed as so affects the variance of the estimates. Whether the result is due to this can only be fully addressed by a simulation study.

3.4. The positive correlation between Ka and Ks is still present (just)

Using the ML method we reported a Ka–Ks correlation, albeit not as profound as found using approximate methods and nearer the neutral expectations. We find using YN00 a significant correlation: Ka = 0.02 + 0.073 Ks; $R^2 = 1.9\%$, $P = 0.005$ (Fig. 4). This is not sensitive to non-parametric analysis (correlation of ranks, $P < 0.0001$) and cannot therefore be due to a few outliers (cf. Makalowski and Boguski, 1998; Smith and Hurst, 1998).

Importantly, the slope (about 0.1) and R^2 for this analysis are very similar to those found in the analysis by Bielawski et al. (2000) and our previous work (Smith and Hurst, 1999b), both of which use ML analysis. The only discrepancy is in the level of significance, a difference which can be accounted for by sample size difference. By contrast the Li based protocol reports a much steeper slope and stronger correlation: Ka = −0.0091 + 0.253 Ks; $R^2 = 12.1\%$, $P < 0.0001$ All the ML and ML approximation analyses agree, therefore, that the covariance is very much weaker than observed using earlier ad hoc methods and is closer to or less than neutral expectations. It is unclear whether a correlation with such a low R^2 value is of great biological significance.

3.5. The rate of protein evolution is lower in GC rich regions

Following Williams and Hurst's (2000) finding that GC and Ka (and Ka/Ks) are negatively correlated, we have

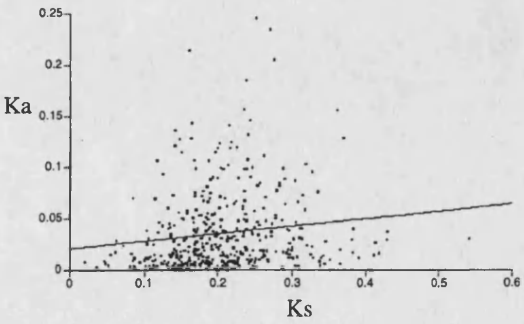


Fig. 4. The relationship between Ka and Ks for 419 autosomal genes in the mouse-rat comparison. Ka and Ks are here calculated using the method of Yang and Nielsen (2000).

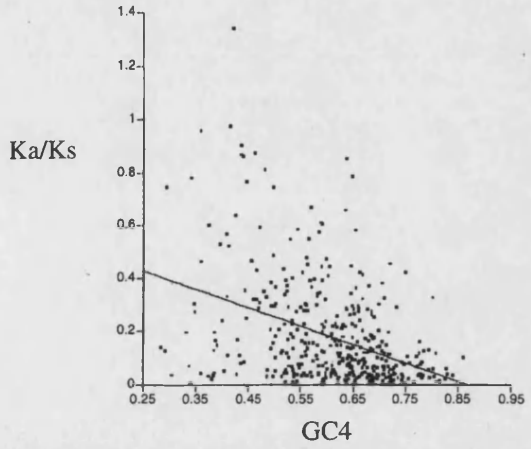


Fig. 5. The relationship between Ka/Ks and GC4 for 419 autosomal genes in the mouse-rat comparison. Ka and Ks are here calculated using the method of Yang and Nielsen (2000).

repeated the analysis using YN00. We find this is a robust finding: Ka = 0.106 − 0.117 GC4; $R^2 = 11.2\%$, $P < 0.0001$: Ka/Ks = 0.603 − 0.696 GC4; $R^2 = 15.4\%$, $P < 0.0001$ (Fig. 5). These concur with the results from the more approximate methods. Using these we find: Ka = 0.102 − 0.109 GC4; $R^2 = 9.5\%$, $P < 0.0001$: Ka/Ks = 0.537 − 0.556 GC4; $R^2 = 10.0\%$, $P < 0.0001$.

4. Discussion

4.1. Is YN00 a reasonable fast method?

The calculation of Ka and Ks can be achieved by very many different methods. Broadly these can be divided into ad hoc methods (the majority) and maximum likelihood methods. We previously reported that the ML method, which also allows for codon usage bias and Ts/Tv bias, recovers at least two results that the earlier ad hoc methods, that make no allowance of codon usage bias, did not recover: a correlation between GC4 and Ks, and only a weak correlation between Ka and Ks.

At the very least we can note that these results are sensitive to methodology. Indeed, in our previous consideration, we noted just this, while not advocating one method to be more robust than another (Smith and Hurst, 1999b). Others might, however, argue that we were too cautious in that (a) in principle one would expect the methods that allow for codon usage bias and Ts/Tv bias should be more accurate than those that do not and (b) the ad hoc estimators are not so easily defended as a maximum likelihood method. However, the issue is not so cut and dried. One issue is whether a pairwise comparison provides enough information for an ML analysis to provide an unbiased estimate (Smith and Hurst, 1999b). Further, it is likely that the ML methods are less accurate when the gene size is small.

If, however, one does suppose that the ML method is the most robust, then clearly this should be the method of choice, as advocated by Yang and Nielsen (Yang and Nielsen, 2000). However, there is an important issue concerning computational time. The ML method is very slow while the most recent of the ad hoc method (YN00) is all but instantaneous. The question to address then is whether the ad hoc approximation to the ML method is adequately accurate.

In this paper we have examined this issue by asking whether YN00 recovers the same pair of exceptional results that the ML method produced. We have shown that it does. We have additionally shown that all methods concur that Ka (and Ka/Ks) correlates negatively with GC4.

Given that YN00 is essentially similar to Li's protocol but with added correction for codon usage bias, we can also now be fairly sure that the discrepancies are owing to this difference. In our previous results, for example, it could have been a distinction between ML and ad hoc estimates that gave rise to the discrepancies and not differences in assumptions.

In summary, we conclude that for analysis of large datasets in which computational time is at a premium, the method of YN00 might represent the optimal balance between computational time and reliability of the results. Given this, it is fair to consider further the results that we have obtained. The weaker Ka–Ks correlation has been discussed previously (Smith and Hurst, 1999b; Bielawski et al., 2000) so we shall not discuss it further. Instead we shall consider the problems in interpreting the apparent correlation of the GC4 and Ks.

4.2. What does the correlation between GC4 and Ks mean for the evolution of isochores?

If one supposes that Yang's methods corrects for previous biases in estimating Ks at extremes of GC, then we find that genes in GC rich regions have the highest silent site substitution rate. While the R^2 value for the linear regression indicates that GC content fails to explain much of the variance in Ks ($R^2 = 9.0\%$), it might also be commented that this is one of the stronger predictors of Ks observed to date. Indeed, there are few other predictors of Ks of autosomal genes. The only other important predictors of Ks on autosomal genes are (1) whether the gene is imprinted or not (Smith and Hurst, 1999a) and (2) whether the genes are linked (Matassi et al., 1999; Williams and Hurst, 2000). The former is a strong effect (imprinted genes have a Ks approximately 2/3 that of other autosomal genes), but very few genes are imprinted. The local similarity of Ks of genes within 1 cM of each other are weak effects: typically R^2 for the regression of Ks for linked genes are about 2% (using the methods that showed no correlation with GC4).

What does this result mean for the study of the evolution of isochores? First it must be noted that any result from the

mouse-rat comparison might not be representative of what happens in isochores in other mammals. While in murids there is a strong local similarity of GC content (Matassi et al., 1999; Williams and Hurst, 2000), the isochore structure in this group is different from that in the rest of the mammals (Robinson et al., 1997; Galtier and Mouchiroud, 1998), showing much less variation in GC.

This is by no means the only problem in interpreting the results. A putative lack of correlation between GC4 and Ks has previously been used as evidence against the DNA repair neutralist models of isochore evolution (Bernardi et al., 1993). The possible inverted u-form of the relationship, consistent with the nucleotide misincorporation/replication time model, has been dismissed as a methodological artefact. If the patterns that we describe are real, it appears that this rejection was valid, but the flat line relationship also appears to be dependent on the methodology.

One might suppose that if the absence of a correlation was evidence against neutralist models (Bernardi et al., 1993), then their rejection on these grounds would appear to be premature, as a correlation is found using what are probably the most robust methods. It is certainly inappropriate to use the claimed lack of a correlation between GC and Ks as definitive evidence either against or for any model. What is more, we have also provided evidence that some of the correlation is probably due to a mutation bias, namely an excess of CpG → TpG mutations in GC rich regions. More thorough investigation of this point will require simulation analyses.

However, there remains a residual correlation even after this bias is accounted for. How can we interpret this? Perhaps our methods missed some of the methyl associated mutations due to multiple hits? We cannot be sure. We can ask whether selectionist or neutralist models might be able to explain the residual pattern observed.

If the nucleotide pool/replication time model really predicts an inverted u-shaped function (Gu and Li, 1994). our results reject this specific model as a u-shaped function is found before and after removal of methyl associated mutations. However, there are positions in parameter space discussed by Gu and Li where the inverted-u or inverted v forms disappear and instead a flatter form, possibly tending to a flat u shaped form, is predicted. This occurs when the so-called next-nucleotide effect is strong. It is possible that the data that we have presented fits such a distribution. However, Gu and Li (1994) argue that empirical data supports a weak next-nucleotide effect. If this is so we can probably reject this model, but this is a weak rejection and is based on the premise that the ML and ML approximation methods are the most reliable.

We have similar problems knowing whether we can reject the repair model (Filipski, 1988; Ticher and Graur, 1989). When a negative correlation between GC and Ks was first observed, it was argued that this was evidence that repair is worse in GC poor regions. We could, naturally turn our finding around, and argue that it could be evidence for better repair in GC poor regions.

4.2.1. The evidence appears to reject biased gene conversion as the explanation for isochores

While we cannot reject most neutralist models, the result is strong evidence against the hypothesis that biased gene conversion is a dominant process. In biased gene conversion one strand from one allele is paired with a strand from the other. Where the two alleles differ these appear as mismatches. These mismatches are then corrected. However, the direction of correction need not be determined at random, in which case the process can end up biased. Importantly, in mammals this process tends to favour GC/AT mismatches becoming GC (Brown and Jiricny, 1988). Such a bias can certainly act to increase the GC level of a sequence and could then be a viable candidate for a force in the evolution of isochores. Such a model would predict a relationship between GC content and recombination rate, which is found (Eyre-Walker, 1993).

However, Eyre-Walker and Bulmer (1995) show (Eq. (9)) that if biased gene conversion is the dominant force, then, through most parameter space we expect a decline in Ks with GC3, although not a dramatic decline until GC3 is very high. An interaction with enhanced recombination rates in GC rich regions would exacerbate the effect. Variation in recombination frequency leads to variation in the frequency of biased gene conversion. Regions with high rates of recombination have high rates of biased gene conversion and, as this process tends to force a GC bias, therefore high G + C. At the same time they are predicted to have low Ks, because new AT ↔ GC mutations in high GC content sequences are most usually GC → AT, and are opposed by biased gene conversion and therefore do not go to fixation. At the extreme, biased gene conversion is so strong that the sequence is all GC and all AT ↔ GC mutations are GC → AT, which are strongly opposed by biased gene conversion. There will then be no substitutions.

It appears hard then to reconcile the present result with this model. However, there may be further complications that prevent firm rejection. Most notably if GC regions have hypermutable sites then the increased Ks could be due to this alone, while at the same time biased gene conversion is acting.

4.2.2. Does selection explain isochores?

This rejection of biased gene conversion is potentially an important rejection. Previous analysis of the pattern of polymorphism at silent sites in mammals suggests that either biased gene conversion or selection might explain the pattern, but not mutation bias (Eyre-Walker, 1999). If this sets of results and ours are robust and reliable then the two together point, by elimination, to selection on silent sites. However, it is unclear that selection can explain the GC4 Ks effect. The same calculations as lead to rejection of biased gene conversion apply equally to directional selection. The pattern expected under stabilising selection is less clear.

There is likely to be variation in the strength of selection around the rodent genome owing to heterogeneity in recom-

bination rates: recombination is most common in GC rich regions (Eyre-Walker, 1993; Williams and Hurst, 2000) and recombination promotes selection (Nordborg et al., 1996). However, while it seems easy to suppose that the low Ka and Ka/Ks found in GC rich regions is consistent with strong stabilising selection (Williams and Hurst, 2000), it is much harder to understand why Ks shows the opposite pattern. In this regard the present finding represents a challenge to selectionist hypotheses for isochore formation. However, rejection of selectionist models will require precise formulation of predictions. For example. in Eyre-Walker and Bulmer (1995), there exist areas in parameter space in which increasing selection can increase the rate of evolution.

4.2.3. Summary

Unfortunately then our data fails to resolve the debate. However, it is nonetheless helpful to know (1) that previous rejection of neutralist models on the basis of a lack of correlation between GC4 and Ks should not be considered a robust dismissal, (2) that the pattern can to some degree probably be explained by a mutation bias, namely, an excess of CpG → TpG mutations in GC rich regions (3) that the pattern is apparently inconsistent with the biased gene conversion model for isochore formation and (4) that while silent site evolution is fastest in GC rich regions, non-synonymous evolution shows the opposite pattern. If we are to believe the results of Yang's modes of estimation, then the first of these results presents a challenge to both neutralists and selectionists.

Acknowledgements

We thank Ziheng Yang for discussion, comments on a previous version and for access to unpublished material. We also thanks three anonymous referees for helpful comments on an earlier version. LDH is funded by the Royal Society.

References

Bernardi, G., Bernardi, G., 1986. Compositional constraints and genome evolution. J. Mol. Evol. 24, 1–11.

Bernardi, G., Mouchiroud, D., Gautier, C., 1993. Silent substitutions in mammalian genomes and their evolutionary implications. J. Mol. Evol. 37, 583–589.

Bernardi, G., Mouchiroud, D., Gautier, C., 1997. Isochores and synonymous substitutions in mammalian genes. In: Bishop, M.J., Rawlings, C.J. (Eds.). DNA and Protein Sequence Analysis, IRL Press, Oxford.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunierrotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Bielawski, J.P., Dunn, K.A., Yang, Z., 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics 156, 1299–1308.

Brown, T.C., Jiricny, J., 1988. Different base base mispairs are corrected with different efficiencies and specificities in monkey kidney-cells. Cell 54, 705–711.

Duret, L., Mouchiroud, D., Gouy, M., 1994. HOVERGEN – a database of homologous vertebrate genes. Nucleic Acid Res. 22, 2360–2365.

Eyre-Walker, A., 1992. Evidence that both G + C rich and G + C poor isochores are replicated early and late in the cell-cycle. Nucleic Acid Res. 20, 1497–1501.

Eyre-Walker, A., 1993. Recombination and mammalian genome evolution. Proc. R. Soc. London Ser. B 252, 237–243.

Eyre-Walker, A., Bulmer, M., 1995. Synonymous substitution rates in enterobacteria. Genetics 140, 1407–1412.

Eyre-Walker, A., 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152, 675–683.

Federico, C., Saccone, S., Bernardi, G., 1998. The gene-richest bands of human chromosomes replicate at the onset of the S-phase. Cytogen. Cell Genet. 80, 83–88.

Filipski, J., 1987. Correlation between molecular clock ticking, codon usage, fidelity of DNA-repair, chromosome-banding and chromatin compactness in germline cells. FEBS Lett. 217, 184–186.

Filipski, J., 1988. Why the rate of silent codon substitution is variable within a vertebrate's genome. J. Theor. Biol. 134, 159–164.

Galtier, N., Mouchiroud, D., 1998. Isochore evolution in mammals: a human-like ancestral structure. Genetics 150, 1577–1584.

Goldman, N., Yang, Z.H., 1994. Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Gu, X., Li, W.H., 1994. A model for the correlation of mutation-rate with GC content and the origin of GC-rich isochores. J. Mol. Evol. 38, 468–475.

Li, W.-H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol. Evol. 36, 96–99.

Li, W.-H., Wu, C.I., Luo, C.C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2, 150–174.

Makalowski, W., Boguski, M.S., 1998. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. J. Mol. Evol. 47, 119–121.

Matassi, G., Sharp, P.M., Gautier, C., 1999. Chromosomal location effects on gene sequence evolution in mammals. Curr. Biol. 9, 786–791.

Mouchiroud, D., Gautier, C., Bernardi, G., 1995. Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J. Mol. Evol. 40, 107–113.

Nordborg, M., Charlesworth, B., Charlesworth, D., 1996. The effect of recombination on background selection. Genet. Res. 67, 159–174.

Pamilo, P., Bianchi, N.O., 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol. Biol. Evol. 10, 271–281.

Robinson, M., Gautier, C., Mouchiroud, D., 1997. Evolution of isochores in rodents. Mol. Biol. Evol. 14, 823–828.

Smith, N.G.C., Hurst, L.D., 1998. Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. 47, 493–500.

Smith, N.G.C., Hurst, L.D., 1999a. The causes of synonymous rate variation in the rodent genome: can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152, 661–673.

Smith, N.G.C., Hurst, L.D., 1999b. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics 153, 1395–1402.

Suoka, N., 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. USA 85, 2653–2657.

Ticher, A., Graur, D., 1989. Nucleic-acid composition. codon usage. and the rate of synonymous substitution in protein-coding genes. J. Mol. Evol. 28, 286–298.

Williams, E.J.B., Hurst, L.D., 2000. The proteins of linked genes evolve at similar rates. Nature 407, 900–903.

74

Wolfe, K.H., Sharp, P.M., 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37, 441–456.
Wolfe, K.H., Sharp, P.M., Li, W.-H., 1989. Mutation rates differ among regions of the mammalian genome. Nature 337, 283–285.
Yang, Z.H., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.

Chapter 6: Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes

Elizabeth J. B. Williams and Laurence D. Hurst (2002)

Journal of Molecular Evolution, 54: **511-518**

J Mol Evol (2002) 54:511–518
DOI: 10.1007/s00239-001-0043-8

JOURNAL OF MOLECULAR EVOLUTION

Clustering of Tissue-Specific Genes Underlies Much of the Similarity in Rates of Protein Evolution of Linked Genes

Elizabeth J.B. Williams, Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

Received: 27 August 2001 / Accepted: 1 October 2001

Abstract. Are genes nonrandomly distributed around the genome and might this explain why it was found that, in the mouse genome, proteins of linked genes evolve at similar rates? Anecdotal evidence suggests that the similarity of expression of linked genes might, in part, explain the similarity in their rates of evolution. Immune system genes, for example, are known to evolve at a high rate and sometimes cluster in the genome. Here we develop methods for statistical tests of similarity of expression of linked genes and report that there is a significant tendency for genes of similar expression breadth to be linked. Significantly, when we exclude tissue specific genes from our sample, the similarity in rates of protein evolution of linked genes is greatly diminished, if not abolished. This diminution is not a sampling artifact. In contrast, while half of the immune genes in our sample reside in 1 of 10 immune clusters in the mouse genome, this clustering appears not to affect the extent of local similarity in rates of evolution. The distribution of placentally expressed genes, in contrast, does have an effect.

Key words: Expression patterns –Linkage –Rate of evolution –MHC

Introduction

Are genomes strings of genes in no particular order or might it be the case that selection favors certain genes to be clustered, possibly to ensure coregulation? While operon structures are well described in bacteria, the linkage of coexpressed genes in eukaryotes is typically considered the exception rather than the rule. However, this view might be changing. In the human genome highly expressed genes appear to be clustered (Caron et al. 2001). Similarly, recent systematic evidence indicates that skeletal muscle genes (Bortoluzzi et al. 1998), extraembryonically expressed genes (Ko et al. 1998), olfactory genes (Lander et al. 2001), and tRNA genes (Lander et al. 2001) tend to show clustering (although only the analysis of extraembryonic genes controls for tandem duplication). Likewise, genes in the MHC cluster tend to be involved in immune functions, and in some cases the most tightly linked (e.g., *Tap* and *LMP*) are involved in coupled processes (Hughes and Yeager 1997).

Here we compile data on expression profiles of a few hundred mouse genes, of known genomic location, to ask whether similarly expressed genes tend to be linked more often than expected by chance. To achieve this we develop measures of similarity of expression. In particular, we examine (1) the breadth of expression, meaning the number of tissues in which a gene is expressed, and (2) the degree of coexpression, meaning the correspondence between genes in the degree to which they are expressed in the same tissues. These two are logically distinct, as two tissue specific genes, for example, will show similar expression breadth but may be expressed in different tissues (i.e., no coexpression). Additionally, we examine a specific coexpression hypothesis. Given that genes in the MHC tend to be immune related, we ask whether

Correspondence to: Laurence D. Hurst; *email:* l.d.hurst@bath.ac.uk

77

immune system genes tend to be clustered more often than expected by chance and whether the MHC might be the exception or the rule.

Expression, Linkage, and Rates of Evolution

The motivation behind these tests is not simply to allow a better statistical appreciation of the degree of ordering in the mouse genome. We also wish to understand whether such patterns might underpin the recently described similarity in the rate of evolution of the proteins of linked genes (Williams and Hurst 2000). For this to be so there needs to be a relationship among expression pattern, linkage, and rates of protein evolution.

Evidence that expression pattern (broadly defined) might be related to the rate of protein evolution comes from a variety of sources. Importantly, proteins of genes expressed in a tissue-specific manner evolve on average twice as fast as those that are ubiquitously expressed (Duret and Mouchiroud 2000). Further, the proteins of certain tissues tend to evolve faster than others. Most notably, immune system genes evolve about twice as fast as nonimmune genes (Hurst and Smith 1999). It is for this reason that we wish to examine the spatial genomic distribution of immune system genes in particular.

Methods

Data Set Compilation

We compiled a data set of mouse and rat orthologues from scrutiny of entries in HOVERGEN (Duret et al. 1994). Genes were accepted as orthologues if, and only if, the mouse and rat sequences had no other nonrodent sequence separating their branches and at least one nonrodent sequence appeared as a sister group. This resulted in a data set of in excess of 500 gene pairs.

Each of the mouse genes was then inspected at LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), using its accession number, to establish mouse chromosomal location. These chromosomal locations are the same as those described at Mouse Genome Informatics (http://www.informatics.jax.org/). Only autosomal genes with a location specified to the centimorgan (cM) were used, because X-linked genes have unusually low rates of evolution (Smith and Hurst 1999). Pairwise Blast under the default settings was used to eliminate tandem duplicates from the data set. Any reported similarity between linked genes led to the elimination of one of the two. This resulted in a data set of 475 autosomal genes. Of these, 289 had at least one neighbor within 1 cM.

Molecular Evolutionary Analysis

The coding sequence was extracted automatically using the annotations in the GenBank entry. DNA alignments were carried out by PILEUP (Wisconsin Package, GCG) using the default settings. The alignments were checked by eye and modified if the alignment was obviously wrong (e.g., translation of aligned sequences gave a nonfunctional protein). Substitution rates were estimated using the method described by Li (1993; Pamilo and Bianchi 1993), applying Kimura's two-parameter method to correct for multiple hits, and by the maximum likelihood method of Goldman and Yang (1994). For each orthologous gene (mouse-rat) we therefore obtained two estimates for the rate, per site, for both nonsynonymous (K_a) and synonymous (K_s) substitutions. We also calculated the rate of protein evolution, controlling for the underlying mutation rate (K_a/K_s). However, we have found that none of the results that we present below are greatly affected by the choice of method. Therefore, for ease of comparison we report only the results using Li's protocol, except where of unusual interest (precise figures for results using the maximum likelihood protocol available on request).

Expression Data

Expression data were assembled from numerous resources. First, all genes were inspected at Unigene (http://www.ncbi.nlm.nih.gov/UniGene/) and the tissues of confirmed expression were noted. These data are based on EST matches of genes and will give only a positive result; negative results are not reported. Additionally, the expression data given at MGI (http://www.informatics.jax.org) was employed. Finally, the original source papers were consulted. If there is disagreement between or within the resources whether a gene is or is not expressed in a certain tissue, we always count the gene as being expressed, under the supposition that a false positive is considerably less likely than a false negative.

From the source papers we could classify some genes as definitely not being expressed in certain tissues (at least at certain times and in certain strains). When a tissue was actively investigated for expression but none was found, we refer to this as the narrow definition of nonexpression. Using this methodology we can, for each gene, score the expression in any given tissue as present, not present (from narrow definition), or 'ho hit" (not a clear positive or negative due to lack of firm data).

Twenty-two tissues were considered. For each gene, we can obtain a score for the total number of tissues in which expression has been reported. This we define as the breadth of expression. While in principle this value might run from 0 to 22 (no expression to ubiquitous expression), we eliminated all those scoring 0, regarding it as evidence that the expression of the gene has yet to be adequately investigated.

Index of Coexpression (ICE)

Not only can we calculate the breadth of expression, but also we can calculate the degree of coexpression for any given pair of genes. This index of coexpression was calculated as follows. If in a given tissue both genes were expressed, or both were not expressed, then the gene pair scores one for that tissue. Expression of one gene and not the other gives a score of −1. This procedure was followed for each of the 22 tissues and a total score was calculated. This total was then divided by the total number of informative tissues to provide an index of coexpression (ICE) that can run from −1 to +1. An ICE value of +1 means perfect coexpression; both genes were expressed in the same tissues and only those tissues. A negative ICE implies antagonistic expression, i.e., where one gene was expressed, the other was not. An ICE value of 0 means coexpression half the time and antagonistic expression the other half. The definition of an "informative tissue" and of "no expression" depends on the precise model that we use. These we now outline.

Models for ICE

We employed three models that differed in their interpretation of the "no hit" category of expression and how this relates to nonexpression. As the data are derived from matches to EST data, it is not the case that no hit simply means no information; it might indicate absence of expression.

Table 1. Summary of the p and r^2 values obtained using the randomization protocols (a) devised by Lercher et al. (2001)[a] and (b) used by Williams and Hurst (2000)[b]

	No. genes	No. comparisons	K_s		K_a/K_s	
			p	r^2	p	r^2
(a)						
Whole data set	289	223	0.0029	7.2%	0.011	10.6%
Without immune genes	243	181	0.053	7.9%	0.12	8.1%
Tissue-specific genes	134	76	0.034	14.0%	0.031	11.9%
Tissue-specific without immune genes	80	51	0.087	13.3%	0.19	10.2%
Without tissue-specific genes	155	87	0.54	0.0%	0.081	5.3%
Tissue-specific without placentally expressed genes	127	67	0.125	9.5%	0.073	9.3%
(b)						
Whole data set	289	196	0.0001	5.8%	0.0001	5.6%
Without immune genes	243	147	0.02	4.1%	0.008	6.2%
Tissue-specific genes	134	61	0.0061	9.6%	0.013	8.1%
Tissue-specific without immune genes	80	38	0.0094	20.5%	0.0083	16.4%
Without tissue-specific genes	155	74	0.34	0.2%	0.1788	2.1%
Tissue-specific without placentally expressed genes	127	55	0.078	5.0%	0.0501	5.8%

[a] These were obtained by comparing each individual gene's K_s and K_a/K_s values with the average of its neighbors. The p value was obtained from randomizations, and the r^2 value from linear correlation.

[b] These were obtained by pairing linked genes using no gene more than twice in total. The p value was obtained from randomization of the mean difference in K values between the pairs of linked genes. The r^2 value was obtained from linear correlation of the K values of the linked genes.

Model 1: No Hit = No Information. At one extreme we can suppose, conservatively, that "no hit" is synonymous with an absence of information. This is reflected in the calculation of the index of coexpression that we calculate for all pairs of linked genes. In this model, informative tissues are those in which expression is either present or confirmed absent for both genes in the pair. If either gene has a no hit, this is treated as an absence of data so is not counted as an informative tissue. When calculating the mean index in any given set of gene pairs, we calculated a mean weighted by the total number of informative tissues.

This method has the problem that it is biased to reporting high ICE values, as most of the information available confirms the presence of expression. An extreme example is that if there were no confirmed lack of expression, all genes would score either 0 for no matches or 1 for at least one confirmed match.

Model 2: No Hit = No Expression. At the other extreme we can suppose that "no hit" means no expression, in which case the number of informative tissues is always 22. This model tends to report high ICE values when the number of no hits is high. Tissues ignored in model 1 because both genes scored no hit, will now return a +1 value to the score. If the sampling of expression data is extensive and EST matches are well reported, then this should, in principle, provide the most reliable information. However, if the sampling is sparse (as must be the case to some extent if some genes have failed to be detected at all or some tissues are not used extensively in EST studies), then this overestimates the degree of coexpression.

Model 3: A Hybrid Model. In our hybrid model we assume that a "no hit" counts as no expression, but only if in the tissue concerned the other gene in the pair has a confirmed expression pattern. This hybrid model attempts to minimize the effect of poor data on the ICE values in model 2. That is, there may be several data points for any given gene pair that score +1 simply because there are many no hit results. If this is due to poor data, rather than a true reflection of expression, this is a problem. In this model the number of informative tissues is 22 minus the number of tissues where both genes have a no hit.

Statistical Analysis

A randomization protocol was used to analyze how similar the expression profiles of linked genes were. To analyze the extent to which linked genes had similar breadths of expression, for each linked pair we calculated the difference in the total number of tissues in which each gene was expressed. This value was calculated for all the linked gene pairs in the data set, and the mean calculated. This mean was compared with means calculated from 10,000 randomizations of the data set. In the randomizations the expression profile of the genes were conserved and the gene position in the genome was randomized. If a gene was used more than once in the original data, it was used more than once in the randomized data set.

We performed a similar procedure for the analysis of the index of coexpression by each of the three models. We calculated a mean (or weighted mean) index of coexpression for the real data. We then randomized the gene positions and calculated a mean index of coexpression for 10,000 randomized sets. These methods allow us to ask how often we would expect by chance the degree of similarity of expression profiles of linked genes which we obtained from the real data set.

For calculating the similarity in rates of evolution we used the method developed by Lercher et. al. (2001) and that employed by Williams and Hurst (2000). The former differs from the previous randomization protocols as it calculates, for each gene, the mean difference between the gene's value (K_s or K_a/K_s) and the mean value of all its neighboring genes within 1 cM. The mean difference calculated from the real data set is then compared to a set of 100,000 random mean values calculated in the same way from randomized data sets. For each test we report (Table 1a) the p value and the r^2 value. The latter is calculated by correlating each individual gene's K value with the mean of its neighbors. In Table 1b we also report the results using the method used by Williams and Hurst (2000) as a comparison. In this method a given gene is compared to its nearest two neighbors (or one neighbor if only one other is within 1 cM). Often, however, the choice of nearest neighbor is arbitrary, as recombination maps place many genes at the same position. The results obtained are sensitive to methodology (Lercher et al. 2001). Given the slightly arbitrary nature of the method

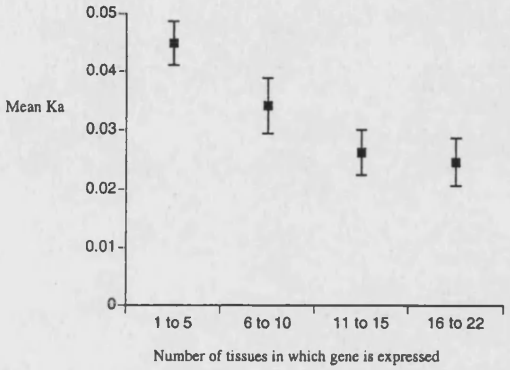


Fig. 1. The relationship between the number of tissues in which a gene is expressed and its rate of nonsynonymous evolution: 1–5 tissues, $N = 133$; 6–10 tissues, $N = 64$; 11–15 tissues, $N = 56$; 16–22 tissues, $N = 32$.

used by Williams and Hurst (2000), we report in the text the results obtained by the method of Lercher et al. (2001) while flagging up results that appear to be method sensitive.

Results

Proteins of Highly Expressed Genes Are Slow-Evolving

We can confirm that within our data set the breadth of gene expression (E) is negatively correlated with the rate of protein evolution ($K_a = 0.046 - 0.0014E$, $r^2 = 3.9\%$, $p < 0.001$) (Fig. 1). Note, however, that this is a relatively weak effect. As with previous analysis, we find no evidence that K_s and expression level covary ($K_s = 0.00018 - 0.0003E$, $r^2 = 0.1\%$, $p = 0.564$).

Proteins of Linked Genes Evolve at a Similar Rate

We can confirm in this data set that the difference in K_a between linked genes is much lower than expected by chance ($p = 0.0029$, $r^2 = 7.2\%$) (Table 1a). Parenthetically, as regards local similarity in K_s, we previously reported (Williams and Hurst 2000) weak significance ($p = 0.01$). In the present data set this effect has decreased marginally ($p = 0.039$, $r^2 = 1.5\%$).

Linked Genes Have a Weak Tendency to Have Similar Expression Patterns

To ask whether linked genes might show similar expression patterns we analyzed the local similarity of expression profiles using two measures.

Expression Breadth. To investigate whether linked genes had similar expression breadths, we calculated the mean difference in breadth of expression (calculated as the total number of tissues in which each gene is expressed) of linked genes and compared this with the mean from 10,000 randomized simulants. We find that only 4% of randomized data sets show a higher level of local similarity in breadth of expression. A priori we would expect that 50% of random data sets would show a higher level of local similarity in expression breadth, therefore this result shows that there is a significant tendency for linked genes to have similar expression breadths.

Degree of Coexpression. Coexpression of linked genes was investigated using the three ICE (index of coexpression) models (explained in methods) for the interpretation of the expression data. Again, we compared the mean (or weighted mean) ICE with the distribution of ICE values obtained through randomization. In each we find at most a weak tendency for linked genes to be more similarly expressed than expected by chance: Model 1 ("no hit" = no information), $p = 0.095$; Model 2 (no hit = no expression), $p = 0.183$; and Model 3 (hybrid model), $p = 0.093$.

Clustering of Immune System Genes Is Very Common. The above results suggest that clusters of genes expressed in the same tissues are the exception rather than the rule. But is this also true if we look more specifically at immune system genes? For these we have a priori expectations that they might be clustered given the presence of the MHC cluster. However, it is hard to provide a definitive definition of what is and what is not an "immune system gene." We chose to apply a method that takes account of as much information as possible. We therefore used all available functional information and expression data and defined a gene as being of the immune system if (a) the knockout had an effect on the immune response or (b) it had expression specific to immune cells (e.g., B cells and T cells). Additionally, Mouse Genome Informatics defines certain genes as belonging to the immune system. We included any gene that MGI considered as belonging to the immune system. No doubt one might query whether our definition is too conservative or too liberal, but in the absence of alternative objective criteria and definitions, we consider this to be a reasonable approach and not obviously prone to bias.

In our data set we find strong evidence for clustering of unrelated immune system genes. There are 46 immune system genes, 24 of which have at least one other immune gene within 1 cM. These exist in 10 clusters, 2 of which are relatively large (Table 2). We could define 13 pairs of linked immune system genes. In 10,000 randomized data sets, on the average there are only 3.75 linked immune pairs (and a maximum of 11). The frequency of

Table 2. The 10 clusters of immune system genes and their chromosomal locations for genes within our sample[a]

Name of gene	Chromosome	cM position	K_a	K_s	K_a/K_s
Interleukin 1 receptor, type I	1	19.5	0.08	0.273	0.293
Interleukin 1 receptor, type II	1	19.5	0.054	0.162	0.333
CD28 antigen	1	30.1	0.055	0.279	0.197
CD152 antigen CTLA	1	30.1	0.046	0.137	0.336
Decay accelerating factor 1	1	67.6	0.185	0.234	0.791
Polymeric immunoglobulin receptor	1	68.5	0.075	0.176	0.426
Cathepsin E	1	69.1	0.036	0.161	0.224
Interleukin 10	1	69.9	0.077	0.173	0.446
Selectin, platelet	1	86.6	0.054	0.228	0.237
CD3 antigen, ζ polypeptide	1	87.2	0.034	0.148	0.230
CD1d1 antigen	3	48.0	0.087	0.21	0.414
CD53 antigen	3	48.5	0.038	0.173	0.220
Small inducible cytokine B subfamily (Cys–X–Cys), mbr 10	5	53.0	0.129	0.329	0.392
Small inducible cytokine B subfamily, mbr 5	5	53.0	0.115	0.241	0.477
CD9 antigen	6	57.0	0.032	0.166	0.193
Tumor necrosis factor receptor superfamily, mbr 1a	6	57.1	0.092	0.184	0.50
Chemokine (C–C) receptor 1, -like 2	9	72.0	0.042	0.143	0.293
Chemokine (C–C) receptor 2	9	72.0	0.031	0.136	0.228
Small inducible cytokine A2	11	46.5	0.098	0.099	0.989
Small inducible cytokine A11	11	47.0	0.025	0.062	0.403
Small inducible cytokine A5	11	47.0	0.023	0.115	0.200
Small inducible cytokine A3	11	47.6	0.064	0.145	0.441
Histocompatibility 2, class II, locus DMa	17	18.56	0.069	0.229	0.301
Tumor necrosis factor	17	19.06	0.035	0.157	0.223

[a] A cluster is defined as the presence of one or more immune system genes within 1 cM of another immune gene. Also listed are the rates of nonsynonymous (K_a) and synonymous evolution (K_s). For the data set as a whole the mean K_a is 0.04 ± 0.04 and the mean K_si s 0.174 ± 0.05. The mean K_a/K_s for these genes is 0.21 ± 0.21, but for these linked immune system genes it is 0.39 ± 0.12.

linked immune system genes is therefore significantly higher than expected by chance ($p < 0.0001$).

Clustering and Rates of Protein Evolution

The above set of results presents highly contrasting pictures: broad-scale analyses report only weak effects, at most, for the effects of linkage on similarity of expression. In contrast, within the same data sets there is a strong pattern of clustering of immune system genes, an effect that is diluted in the broad-scale pattern. A priori given the weakness of the broad-scale patterns, it seems unlikely that a broad-scale analysis of the extent to which expression similarity covaries with the similarity of rates of evolution of linked genes will provide an informative result. However, this seems not to be the case.

Clustering of Tissue-Specific Genes Underlies the Local Similarity in Rates of Evolution. A gene can be defined as being tissue specific if it is expressed in fewer than six tissues. This, naturally, is an arbitrary divide. However, genes expressed in only one tissue are too few to provide meaningful analysis. We can then partition our sample into tissue-specific ($N = 134$) and nonspecific ($N = 155$) genes. Within the nonspecific group we find that the local similarity is removed completely (K_a: $p = 0.54$, $r^2 = 0.0\%$) (Table 1a, Fig. 2). In contrast, when we look at tissue-specific genes and test for local similarity in rates of evolution, we find that there is a correlation stronger than before, even though the p value does not indicate that it is highly statistically significant (K_a: $p = 0.034$, $r^2 = 14.0\%$) (Table 1a, Fig. 2). Using the method of Williams and Hurst (2000), the relevant p value resolves to 0.0061 ($r^2 = 9.6\%$), versus $p = 0.0001$ ($r^2 = 5.8\%$) for the data set as a whole. This suggests that the local similarity in the rate of protein evolution is due largely to the distribution and rate of evolution of tissue-specific genes.

Could the apparent absence of local similarity in the non-tissue-specific set of genes be an artifact of dividing the data set up, thereby reducing the sample size? To examine this, we randomly divided the entire data set into two subsamples, one the same size as the tissue-specific group ($N = 134$) and the other containing the remainder ($N = 155$). We repeated this 100 times. We then calculated the extent of local similarity in each random subsample using the method of Lercher et al. We found that in none of the samples did the larger half give

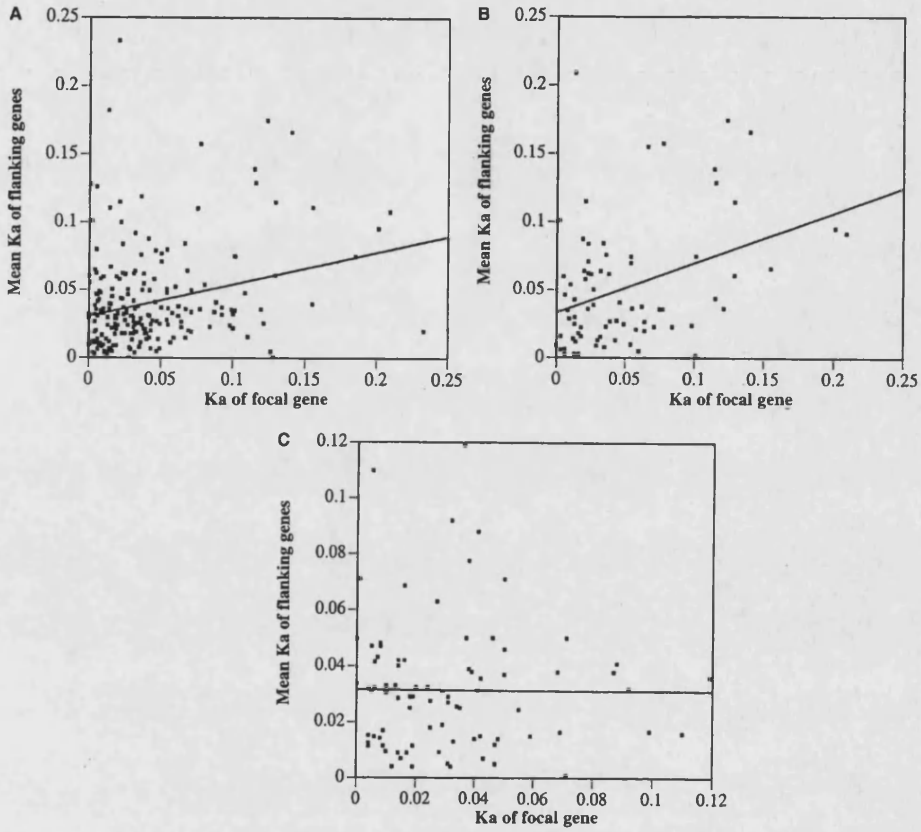


Fig. 2. The relationship between the K_a of a focal gene and the mean K_a of the surrounding genes for **(A)** the complete sample, **(B)** the tissue-specific genes, and **(C)** the non-tissue-specific genes.

an r^2 near 0.0% or the small half give such a high r^2 value. This indicates that this result is not an artifact of subsampling per se. As regards K_a/K_s, the r^2 in the non-tissue-specific sample is approximately half that in the complete data set. In none of 100 trials did this amount of diminution occur.

What classes of tissue-specific genes are there that could possibly be responsible for this effect? There is an a priori expectation that genes involved in antagonistic coevolution may evolve at high rates. If these too are clustered, then this will lead to local similarity in the rates of evolution. This is because in randomized data sets on the average, these fast-evolving genes tend to be paired up with slower-evolving genes. They would thus cause the randomized data sets to have a higher average difference in K_a between linked genes than in the real data set. We have shown that immune genes tend to cluster and it is well established that they have unusually high rates of evolution, probably because of host parasite coevolution. Similarly, genes putatively involved in mother–offspring conflict may show increased rates of evolution (Hurst and McVean 1998). Many of these are likely to be placentally expressed. This is of significance, as prior evidence suggests that placentally expressed genes are clustered (Ko et al. 1998). We therefore examined the consequences of removal of immune and placental genes. Given the absence of a priori expectations for other sets of genes for which we have data, we shall not examine any other subcategories.

Removal of the immune system genes has a little effect on the extent of local similarity as assayed by the r^2 values. Now using 243 genes (i.e., the complete set minus the immune genes), we find a comparable amount of local similarity in K_a as in the complete data set (K_a, $r^2 = 7.9\%$ and $p = 0.05$; K_a/K_s, $r^2 = 8.1\%$ and $p = 0.12$) (Table 1a). When we examine 100 random data sets, each containing 243 randomly selected entries for the original data set, we find that the r^2 value from the nonimmune data set is not unusual. Indeed, in the case of K_a, the r^2 increases. This indicates that the clustering of immune genes is not of importance in determining the local similarity in rates of evolution.

Given the lack of effect on the r^2 values, the decline in the p value seen on the removal of immune genes most

likely reflects sample size changes. This we confirmed. We took each of the 100 randomly assembled data sets of 243 genes and measured the mean local similarity within each using the method of Lercher et al. Then we did 10,000 randomizations of each of these 100. We then asked what proportion showed a greater mean local similarity and thereby determined a p value for each of the 100 sets. We found that 12% of the random collections reported a p value above that shown in the nonimmune data set. We therefore failed to reject the hypothesis that the weakening of the p value in the nonimmune set is anything other than a sampling effect.

These conclusions are further supported by analysis of the tissue-specific genes. Within the tissue-specific group without the immune genes, the local similarity is increased (from $r^2 = 9.6\%$ in the complete set of tissue-specific genes to $r^2 = 20.5\%$ after the removal of immune genes) under the protocol of Williams and Hurst (2000). Under the protocol of Lercher et al (2001), the r^2 remains largely unchanged ($r^2 = 14$ versus 13.3%).

The distribution of seven placentally expressed genes, in contrast, appears to have an effect on the local similarity within the class of tissue-specific genes. When these are removed from the tissue-specific gene class, the local similarity decreases and is not statistically significant under either model (method of Lercher et al.—K_a, $p = 0.125$ and $r^2 = 9.5\%$; method of Williams and Hurst—K_a, $p = 0.078$ and $r^2 = 5.0\%$). Again using the method of randomly subsampling, this time randomly removing seven genes from the tissue-specific data set, none of the 100 random subsamples showed such dramatic decreases in r^2.

Discussion

In this paper we have set out to ask two questions. First, do similarly expressed genes tend to cluster in the genome? Second, if they do, does this explain why linked genes evolve at similar rates? We have found evidence that there is a significant tendency for genes of comparable expression breadth to be linked but only a weak tendency, at most, for genes that are coexpressed to be linked. One limitation of our study is the usage of expression data that permit us to analyze presence or absence of expression rather than rate of expression, which might be the more relevant parameter. In the near-future results from microarray data and SAGE analyses should allow exploration of these issues as well.

Given the weakness of the tendency for genes of comparable expression to be linked, and the weakness of the correlation between K_a and expression breadth, we might reasonably conclude that it is a priori unlikely that linkage of similarly expressed genes might explain why linked genes evolve at similar rates. This, however, appears not to be so: within the class of nonspecific genes there is no tendency for linked genes to have similar rates of protein evolution. The local similarity in rates of evolution appears to be due in no small part to the genomic positioning of tissue-specific genes. This is due in part to clustering of placentally expressed genes but is not dependent on the clustering of immune system genes.

These results do not examine whether coexpression more generally underlies local similarity in rates of evolution. Unfortunately, here we can perform only much weaker tests. In yeast, the member of a pair of duplicates that has the higher expression level has the higher rate of protein evolution (Pal et al. 2001b). Evidence that this is so came from analysis of the regression of the difference in the rate of protein evolution versus the difference in expression level (assayed by microarray data) for each pair of duplicates. We can attempt the same sort of analysis for the present data set. That is, if similarity of expression pattern does explain some of the local similarity of rates of protein evolution, then we expect that a large local difference in K_a should reflect a large difference in expression profile.

To see whether this occurs we can plot ΔK_a (pairwise difference in K_a) versus ICE for each pair of linked genes. If coexpression predicts the local similarity in K_a to any extent, then we expect a negative correlation between ΔK_a and ICE. We do not find this: ΔK_a versus ICE, Model 1 ($\Delta K_a = 0.03 - 0.001$ ICE1; $r^2 = 0.001\%$, $p = 0.61$); ICE, Model 2 ($\Delta K_a = 0.03 - 0.005$ ICE2; $r^2 = 0.006\%$, $p = 0.295$; and ICE, Model 3 ($\Delta K_a = 0.03 - 0.006$ ICE3; $r^2 = 0.009\%$, $p = 0.175$). However, while we know that the expression breadth covaries with K_a, ΔK_a does not covary with ΔE ($\Delta K_a = 0.0324 - 0.0001\ \Delta E$; $r^2 = 0.00\%$, $p = 0.8$). The latter result indicates that these are very weak tests. The above result must therefore be considered a rejection of the possibility that there is a strong covariation of expression and rate of evolution. We cannot therefore make any strong conclusions regarding coexpression.

The GC K_a Problem

It is remarkable that removal of the tissue-specific genes from the data set destroys the signal of local similarity in rates of protein evolution. This suggests that the effects are unlikely to be genome-wide. This, however, leaves the problem of the causes of the negative correlation between GC content and K_a. Unlike the K_s/GC and K_a/K_s correlations, the K_a/GC correlation is not sensitive to method: the GY94 protocol reports the same result as Li93 (Li93, $K_a = 0.108218 - 0.118808$GC4, $r^2 = 13.1\%$, $p < 0.0001$; GY94, K_a_ML $= 0.122907 - 0.141966$GC4, $r^2 = 16.4\%$, $p < 0.0001$). This negative correlation was considered by Williams and Hurst (2000) to be consistent with the idea that local similarity in rates of protein evolution was due to genome-wide variation in the strength of purifying selection owing to variation in

the recombination rate around the genome (i.e., the intensity of Hill–Robertson effects). This interpretation rests on the understanding that the recombination rate covaries with the GC content (Fullerton et al. 2001). The Hill–Robertson model is given some support by the finding that variation in K_a and K_a/K_s within the *Drosophila* genome covaries negatively with the recombination rate (Comeron and Kreitman 2000).

If the local similarity in rates of protein evolution is due largely to linkage of similarly expressed genes, and disappears when tissue-specific genes are removed, how are we to interpret this strong GC/K_a correlation? One possibility is that, as in yeast, the recombination rate (hence GC) and gene expression rates covary, so a correlation between recombination/GC and K_an eed not be evidence for Hill–Robertson effects (Pal et al. 2001a). We cannot analyze expression rates in mammals. We find, however, that there is a positive correlation between breadth of expression and GC content at fourfold redundant sites ($E = 4.28 + 4.27GC4$, $r^2 = 1.3\%$, $p = 0.06$). Given that broadly expressed genes have low rates of evolution, this is in the right direction to explain why genes with a high GC content might have low rates of protein evolution. The correlation is, however, weak [and, incidentally, in the direction opposite to that reported by Goncalves et al. (2000) for human sequences]. This effect is so weak that it cannot account for the greatly reduced K_a in regions of high GC content. This is confirmed by the finding that the K_a/GC correlation remains when only the tissue-specific genes are analyzed ($K_a = 0.15 - 0.18GC4$, $r^2 = 22.1\%$, $p < 0.0001$, $N = 126$).

Alternatively, it might simply be the case that immune system genes (under directional selection or subject to overdominance) tend to be AT rich. Were this so, the GC/K_a correlation need not be indicative of variation in purifying selection. Indeed, when we divided our data set into immune and nonimmune genes, immune system genes tended to be AT rich (GC4 immune $= 0.55 \pm 0.016$; GC4 nonimmune $= 0.61 \pm 0.008$). However, both sets still showed a strong K_a/GC4 correlation (nonimmune, $K_a = 0.075 - 0.075GC4$, $r^2 = 8.6\%$, $p < 0.001$; immune, $K_a = 0.21 - 0.24GC4$, $r^2 = 24.1\%$, $p = 0.001$).

Given that these two possible explanations do not explain the GC/K_a correlation, we must regard the cause as problematic. Given that the result is both relatively strong and robust to methodology (unlike the K_a/K_s correlation), the causes of the correlation deserve further scrutiny.

Acknowledgments. We thank Deborah Charlesworth, Martin Lercher, and Araxi Urratia for their constructive comments.

References

Bortoluzzi S, Rampoldi L, Simionati B, Zimbello R, Barbon A, d'Alessi F, Tiso N, Pallavicini A, Toppo S, Cannata N, Valle G, Lanfranchi C, Danieli GA (1998) A comprehensive, high-resolution genomic transcript map of human skeletal muscle. Genome Res 8:817–825

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R (2001) The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291:1289–1292

Comeron JM, Kreitman M (2000) The correlation between intron length and Recombination in *Drosophila*: Dynamic equilibrium between mutational and selective forces. Genetics 156:1175–1190

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17:68–74

Duret L, Mouchiroud D, Gouy M (1994) Hovergen—A database of homologous vertebrate genes. Nucleic Acids Res 22:2360–2365

Fullerton SM, Carvalho AB, Clark AG (2001) Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18:1139–1142

Goldman N, Yang ZH (1994) Codon-based model of nucleotide substitution for protein-coding dna sequences. Mol Biol Evol 11:725–736

Goncalves I, Duret L, Mouchiroud D (2000) Nature and structure of human genes that generate retropseudogenes. Genome Res 10:672–678

Hughes AL, Yeager M (1997) Molecular evolution of the vertebrate immune system. Bioessays 19:777–786

Hurst LD, McVean GT (1998) Do we understand the evolution of genomic imprinting? Curr Opin Genet Dev 8:701–708

Hurst LD, Smith NGC (1999) Do essential genes evolve slowly? Curr Biol 9:747–750

Ko MSH, Threat TA, Wang XQ, Horton JH, Cui YS, Wang XH, Pryor E, Paris J, WellsSmith J, Kitchen JR, Rowe LB, Eppig J, Satoh T, Brant L, Fujiwara H, Yotsumoto S, Nakashima H (1998) Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome. Hum Mol Genet 7: 1967–1978

Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lercher MJ, Williams EJB, Hurst LD (2001) Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons. Mol Biol Evol 18:2032–2039

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Pal C, Bapp B, Hurst LD (2001a) Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. Mol Biol Evol 18:2323–2326

Pal C, Bapp B, Hurst LD (2001b) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931

Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281

Smith NGC, Hurst LD (1999) The causes of synonymous rate variation in the rodent genome: Can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152:661–673

Williams EJB, Hurst LD (2000) The proteins of linked genes evolve at similar rates. Nature 407:900–903

84

Chapter 7: The molecular evolution of signal peptides (2000)

Elizabeth J. B. Williams, Csaba Pal and Laurence D. Hurst

Gene, 253: **313-322**

ELSEVIER

Gene 253 (2000) 313–322

GENE
AN INTERNATIONAL JOURNAL ON GENES, GENOMES AND EVOLUTION

www.elsevier.com/locate/gene

The molecular evolution of signal peptides

Elizabeth J.B. Williams, Csaba Pal, Laurence D. Hurst *

Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

Received 3 March 2000; received in revised form 12 May 2000; accepted 24 May 2000
Received by W.-H. Li

Abstract

Signal peptides direct mature peptides to their appropriate cellular location, after which they are cleaved off. Very many random alternatives can serve the same function. Of all coding sequences, therefore, signal peptides might come closest to being neutrally evolving. Here we consider this issue by examining the molecular evolution of 76 mouse–rat orthologues, each with defined signal peptides. Although they do evolve rapidly, they evolve about half as fast as neutral sequences. This indicates that a substantial proportion of mutations must be under stabilizing selection. A few putative signal sequences lack a hydrophobic core and these tend to be more slowly evolving than others, indicating even stronger stabilizing selection. However, closer scrutiny suggests that some of these represent mis-annotations in GenBank. It is also likely that some of the substitutions are not neutral. We find, for example, that the rate of protein evolution correlates with that of the mature peptide. This may be a result of compensatory evolution. We also find that signal peptides of immune genes tend to be faster evolving than the average, which suggests an association with antagonistic co-evolution. Previous reports also indicated that the signal peptide of the imprinted gene, *Igf2r*, is also unusually fast evolving. This, it was hypothesized, might also be indicative of antagonistic co-evolution. Comparison of *Igf2r*'s signal peptide evolution shows that, although it is not an outlier, its rate of evolution is comparable to that of many of the faster evolving immune system signal sequences and 5/6 of the amino acid changes do not conserve hydrophobicity. This is at least suggestive that there is something unusual about *Igf2r*'s signal sequence.

Keywords: *Igf2r*; Molecular evolution; Mouse–rat orthologues; Neutral evolution; Signal peptides

1. Introduction

There is considerable variation both within and between proteins in the rate of evolution. What are the causes of this variation? Most proteins in the mouse–rat comparison (Wolfe and Sharp, 1993; Makalowski and Boguski, 1998; Hurst and Smith, 1999) show evidence of stabilizing selection and hence evolve at a slow rate, compared with the underlying mutation rate (as assayed by the K_A/K_S ratio, where K_A and K_S are the rates of non-synonymous and synonymous DNA changes per site, respectively). However, there remain a significant proportion of proteins (or sub-domains of proteins) that show relatively high rates of evolution.

Numerous analyses have indicated that many fast evolving proteins, or sub-domains of proteins, are probably engaged in some form of selectively driven antagonistic co-evolution. For example, host immune system genes (Hughes et al., 1990; Hurst and Smith, 1999) and parasite antigens (Hughes, 1992) show rapid evolution, especially at the sites of mutual binding (Hughes et al., 1990; Hughes, 1991, 1992). Similarly, genes involved in bacterial antagonistic interactions (Tan and Riley, 1997), as well as those potentially involved in both inter-sexual conflict, for example some of those of *Drosophila*'s seminal fluids (Aguade et al., 1992; Tsaur and Wu, 1997), and parent–offspring conflict, for example numerous placentally expressed genes in mammals (e.g., placental lactogens) (Hurst and McVean, 1998), also show unusually high rates of evolution.

However, other high rates of evolution are indicative of neutral evolution. The most convincing example comes from analysis of pseudogenes (Li et al., 1981). However, there has yet to be identified a class of protein coding genes (or sub-domains) that are dominantly

Abbreviations: lgf2r, insulin like growth factor type II receptor; lgf2, insulin like growth factor 2; Q, glutamine; L, leucine; P, proline; R, arginine; V, valine.

* Corresponding author. Tel.: +44-1225-826424; fax: +44-1225-826779.
E-mail address: l.d.hurst@bath.ac.uk (L.D. Hurst)

0378-1119/00/$ - see front matter
PII: S0378-1119(00)00233-X

96

neutrally evolving. Here we examine the molecular evolution of signal peptides and ask whether these might serve as good paradigmatic examples of neutral evolution, as from knowledge of their biochemistry this might be suspected.

Signal peptides are short N terminal (genic 5′) parts of a protein whose function it is to direct the peptide to its appropriate cellular location. After having delivered the mature peptide to this location, the signal peptide is cleaved and is presumed to be digested. The fact that it is cleaved allows us to suppose that signal peptides have one and only one function, to deliver the mature peptide to its appropriate location.

Signal peptides often show little evidence of sequence similarity. The lack of identity among these sequences implies that numerous forms of sequence can serve the very same role and are sufficient for membrane transport (see, e.g., Izard and Kendall, 1994). It has often been argued that the only requirement for proper functioning of the signal peptide is to contain a hydrophobic core consisting exclusively or largely of hydrophobic amino acids. As a support for the theory, Kaiser et al. (1987) found that about 20% of random sequences can act as functional signal sequences. Furthermore, it is also known that amino acids with similar hydrophobicity are coded by neighbouring codons (see for references Freeland and Hurst, 1998). Therefore, most non-synonymous mutations conserve hydrophobicity. More generally then, if signal sequences are not neutrally evolving, it is hard to imagine a class of coding sequences (as opposed to pseudogenes) that are wholly neutral (but see Dickerson, 1971 on fibrinopeptides).

There are, however, other features of signal peptides that are common, although the relationship between structure and function is not transparent. Often there is a leucine-rich region. Typical signal peptides also have a positively charged n-region and a neutral but polar c-region. Positions -3 and -1 from the cleavage site must be small and neutral for cleavage to occur correctly (for analysis of other diagnostic features/methods see Nielsen et al., 1997; Ladunga, 1999).

While the previously reported low sequence similarity is consistent with neutral evolution, other data suggest that the pattern might be more complicated. From a sample size of three, it was reported that some genes may have relatively low rates of evolution in the signal sequence (Li et al., 1985), but that this may be an artifact of the small size of the signal peptide. Further, Smith and Hurst (1998) reported a strong correlation between the rate of evolution of the signal peptide and that of the complete gene. This is hard to understand from a neutralist perspective. However, this previous analysis permitted non-independence, in that the signal peptide was allowed as part of the complete coding sequence. So it remains to be established whether, if one controls for non-independence, the correlation of rates still exists.

Here, then, we shall ask just how fast signal sequences evolve and whether their evolution is simply neutral. To this end we shall also ask whether the rate of evolution in the signal peptide is correlated with that in the mature peptide. Additionally, we shall ask whether there are any biochemical aspects of signal peptides that in any way explain the variance in their rate of evolution.

We shall also examine a specific selectionist explanation for the evolution of one particular signal peptide, that of the insulin like growth factor type II receptor (*Igf2r*). *Igf2r* is an imprinted gene expressed off the maternally derived chromosome in rodent embryos. It is a transmembrane receptor that binds to the paternally expressed Igf2 (and numerous other ligands) to target them to the lysosomes. This is interpreted by the conflict hypothesis for the evolution of imprinting (Moore and Haig, 1991) as an antagonistic interaction. McVean and Hurst (1997) found that, in the mouse–rat comparison, at the position where Igf2r binds to Igf2 the protein shows an especially low rate of evolution, indicating stabilizing selection, rather than the directional selection expected if the interaction is antagonistic. Smith and Hurst (1998) showed that the same was also true in the human–cow comparison.

The latter analysis, however, also reported that in both the human–cow and the mouse–rat comparisons, the signal sequence has a high K_A/K_S ratio. It was speculated, therefore, that there might be a conflict concerning the cellular localization of Igf2r. Hence, the unusual property of Igf2r's signal sequence may be explained by strong directional selection driven by antagonistic co-evolution. As the previous analysis used only nine other genes, here we aim to ask whether Igf2r's signal peptide really is an outlier by comparing it with a much larger set of other signal peptides.

2. Materials and methods

We compiled a dataset of mouse and rat orthologues in which the signal peptide has been annotated in at least one of the two GenBank entries. NCBI Entrez (http://www.ncbi.nlm.nih.gov/Entrez/), and ACNUC software at the UK HGMP Resource Centre (http://www.hgmp.mrc.ac.uk/) was used to search and extract rat and mouse complete coding sequences with annotated signal peptide regions. This resulted in a list of nearly 400 genes.

Each of these was scrutinized in the HOVERGEN database (Duret et al., 1994) to find mouse–rat orthologues. Genes were accepted as othologues if, and only if, the mouse–rat sequences had no other non-rodent sequence between them and at least one non-rodent sequence appeared as a sister group. This resulted in a data set of 80 gene pairs.

GENETRANS was used to automatically extract

87

complete coding sequences, while GBPARSE (available from http://sunflower.bio.indiana.edu/~wfischer/Perl_Scripts/) was used to automatically extract signal peptide regions, using annotations in the GenBank entries. Mature peptides (complete sequence minus signal peptide) were analysed by editing out the signal peptide from the alignment files.

DNA alignments of signal sequence and entire protein coding sequences were carried out by PILEUP, using the default settings. The alignments were checked by eye and modified if necessary. Signal sequences were checked to ensure that they aligned perfectly against themselves within the complete gene alignment. For four genes we were unable to find unambiguous alignments, and these were excluded. This resulted in a dataset of 76 genes, including *Igf2r*. The genes and their Accession Nos. are given in Appendix A.

Substitution rates were estimated by the package DIVERGE (available at HGMP). The program is based on the method described by Li (1993) using the Pamilo and Bianchi modifications (Pamilo and Bianchi, 1993), and applies Kimura's two-parameter method to correct for multiple hits and to account for the difference in substitution rates for transitions and transversions. These data are also reported in Appendix A.

The hydrophobicity of all the signal peptides was examined using PepPlot within GCG at HGMP. Mouse Genome Informatics (http://mgd.hgmp.mrc.ac.uk/) and SWISS-PROT were used to find immune related genes of our sample. A gene was classified as an immune gene if either of the two entries specifically mentioned involvement in immune response or expression in immune specific cell types.

3. Results

3.1. Signal peptides have a fast rate of evolution, but many non-synonymous mutations are under stabilizing selection

Signal peptides do appear to evolve faster than mature peptides, although by how much depends precisely on the statistic used (see Table 1). If we calculate a mean K_A for both mature and signal peptide, then we find that on average signal peptides evolve a little under twice as fast. Allowing for underlying mutation rate differences by using the K_A/K_S ratio suggests that signal peptides evolve a little over twice as fast. By comparison, if we consider the mean value of the paired ratios per gene (e.g., K_A signal peptide/K_A mature peptide), then the signal peptides on the average evolve over five times faster than the flanking mature peptide. Allowing for the underlying mutation rate, we find a comparable figure. The paired test is possibly the least accurate as the ratios have an extremely high variance which most likely reflects the effects of the small size of the signal peptide. Similarly, if K_S is unusually low, the K_A/K_S ratio becomes extraordinarily (and probably unrealistically) high. That signal peptides are fast evolving, none the less seems clear: out of 76 genes 53 had a higher K_A/K_S in the signal sequence than in the gene as a whole. This is highly significantly different from null expectations (χ^2, $P<0.001$).

However, by none of these measures is the rate of signal sequence evolution as high as would be expected were the sequences neutrally evolving. Mean K_A is half the value of mean K_S, and mean K_A/K_S for each signal sequence is 0.63 ± 0.114 which is very much more than two standard errors away from unity, the figure expected if sequence evolution is perfectly neutral.

3.2. Unusual signal sequences evolve unusually slowly

Some signal peptides appear to evolve relatively slowly. Is this chance variation or might these sequences be functionally unusual as well? We have examined the hydrophobicity plot of all of the signal peptides. At least six of our genes do not have the typically hydrophobic signal peptide, i.e. they lacked a hydrophobic core (NB. Igf2r is normal). These are indicated in Appendix A. Although the sample size is limited, we find that these six tend to evolve slowly for signal peptides (they evolve at about a third of the rate of others), although the statistics are marginal (Table 2).

Table 1
Basic statistics of the K_A, $K_S$and K_A/K_S for mature and signal peptides ($N=76$)

	K_A	K_S	K_A/K_S
Mature peptide (mean ±S.E.M.)	0.05±0.006	0.198±0.010	0.249±0.028
Signal peptide (mean ±S.E.M.)	0.09±0.012	0.181±0.019	0.628±0.114[b]
Signal/mature (paired)	5.23±1.60[a]	0.990±0.108	5.61±1.08[c]
Rank correlation between mature and signal, *P* value from slope of regression of ranks	$r^2=0.136$	$r^2=0.023$	$r^2=0.033$
	$P=0.001$	$P=0.193$	$P=0.12$
	Slope=0.368	Slope=0.151	Slope=0.18

[a] Omits four data points in which mature $K_A=0$.
[b] Omits three data points in which $K_S=0$.
[c] Omits five data points in which the ratio is infinite.

88

Table 2
Comparison of the six unusual signal peptides with those with normal hydrophobicity plots

	K_A	K_S	K_A/K_S	Signal peptide size (nt)
Unusual ($N=6$)	0.032±0.014	0.123±0.045	0.219±0.065	310±161
Normal ($N=70$)	0.096±0.012	0.186±0.021	0.665±0.123[a]	76±3.39
Mann–Whitney U test for difference	$P=0.0506$	$P=0.3808$	$P=0.066$	$P=0.365$

[a] Omits three data points in which $K_S=0$.

This finding obviously tempts the question as to whether there are different classes of signal peptide that have different rates of evolution (and if so why) or whether these six do not really have signal peptides at all and are mis-annotated in the GenBank entry?

To address this issue further, we examined the Swiss-Prot entries for these six proteins. We have also examined the sequences using Sigcleave at EMBOSS (http://www.hgmp.mrc.ac.uk/Registered/Option/emboss.html). Two of these, Acetyl Co-A (Swiss-Prot Acc: P45952, mouse; P08503, rat) and sterol carrier protein 2 (Swiss-Prot Acc: P32020, mouse; P11915, rat) had no signal peptide mentioned in Swiss-Prot. Sigcleave failed to identify any signal peptide cleavage sites. Sigcleave correctly identifies 95% of signal peptides, and rejects 95% of non-signal peptides. The cleavage site should be correctly predicted in 75–80% of cases. Given this, the GenBank annotation is likely to be misleading.

Of the remaining four, both Sigcleave and Swiss-Prot agreed that a signal peptides might be present. However, Ephrin B1 (Swiss-Prot Acc: P52795, mouse; P52796, rat) and Coagulation factor III (Swiss-Prot Acc: P20352, mouse; P42533, rat) have only weakly defined cleavage sites under the Sigcleave analysis. Ephrin B1 also lacks the usual leucine-rich domain. It is therefore possible that these proteins do not have signal peptides.

Inhibin beta A (Swiss-Prot Acc: P18331, mouse; Q04998, rat) has an unusual Swiss-Prot entry, as the rat protein had been annotated as having a signal peptide and a propeptide; however, it was not known where one stopped and the other started. The GenBank entry may well then be a combination of signal peptide and propeptide. This was supported by Sigcleave analysis. This method found a cleavage site at only 21 amino acids, where the GenBank annotation indicates a signal peptide in excess of 200 amino acids long. The size defined by Sigcleave is around the mean for the remaining 'normal' signal peptides. Inhibin alpha (Swiss-Prot Acc: Q04997, mouse; P17490, rat) likewise has a huge signal peptide according to GenBank, but both Swiss-Prot and Sigcleave agree that the signal peptide is cleaved at amino acid 21. This could have caused the appearance of a slow rate of evolution. However, inhibin beta A shows neither synonymous nor non-synonymous evolution ($K_A=K_S=0$). Inhibin alpha shows a high K_A/K_S ratio ($K_A=0.07$, $K_S=0.04$).

It appears, then, that there is some degree of mis-annotation in GenBank. This issue can, however, only be addressed definitively by detailed biochemical analysis of the genes concerned, analysis which, as yet, appears not to have been done.

3.3. *Mitochondrial signal sequences are longer but evolve at a normal rate*

Signal sequences are known to direct the transport of proteins across different types of membranes (e.g., endoplasmic reticulum, Golgi-network, mitochondria). Therefore, it is reasonable to ask whether the variation in the rate of evolution is explained by the location to which the signal peptides direct the protein. In order to address this issue, we have compared the evolution of signal sequences that direct the import of mitochondrial proteins encoded in the nucleus to the remaining others (Table 3).

In our original sample there is only one sequence that was annotated as being a nuclear-encoded mitochondrial protein. Therefore, we compiled a new dataset of mitochondrial proteins. This we did by examining NCBI Entrez using 'mitochondrial' as key word and

Table 3
Comparison of mitochondrial and non-mitochondrial signal peptides

	K_A sig	$K_S$sig	K_A/K_Ssig	Signal peptide size (nt)
Mitochondrial ($N=8$)	0.0572±0.016	0.102±0.015	0.727±0.251	107.6±12
Non-mitochondrial ($N=75$)	0.089±0.012	0.18±0.02	0.608±0.114[a]	94.2±14.1
Mann–Whitney U test for difference	$P=0.44$	$P=0.19$	$P=0.56$	$P=0.0029$

[a] Omits three data points in which $K_S=0$.

then checking to ensure the genes were nuclear. Although several mitochondrial genes with annotated signal (or transit) peptides can be found in the databank, only eight genes have been found with rat orthologues (see Appendix B). The analyses of these genes have detected no significant difference in the rate of evolution of mitochondrial signal sequences compared with that of non-mitochondrial ones. However, we have to emphasize that the failure to notice any differences may be a pitfall of the low sample size of mitochondrial proteins.

Although there is no sign of unusual evolution, it is still possible that mitochondrial signal peptides are functionally different. We find that mitochondrial genes are significantly longer than non-mitochondrial ones (Mann–Whitney test, $P<0.01$). This result is not surprising, as precursor proteins are imported into the mitochondria in a multistep process mediated by translocation systems of the outer and inner membrane (Gillham, 1995). Hence, pre-sequences of mitochondrial proteins are expected to contain multiple signal elements to reach their appropriate locations (Gillham, 1995). We have also examined the secondary structure of mitochondrial pre-sequences. None of them show signs of unusual hydrophobicity.

3.4. Are the substitutions due to selection or drift?

While signal sequences as a whole are not perfectly neutrally evolving, we can also ask about the substitutions that are seen. Are these the result of drift or might positive selection be suspected? We cannot answer this question definitively, but can ask whether (a) the substitutions greatly affect hydrophobicity, (b) whether the rates of evolution of signal and mature peptides are correlated (which is not obviously consistent with neutral expectations) and (c) whether genes involved in antagonistic interactions (immune genes and Igf2r) show fast evolving signal peptides.

3.5. Is hydrophobicity conserved?

A neutralist model for the evolution of signal sequences would predict that the non-synonymous substitutions conserve the hydrophobicity of the amino acid. This test does not discriminate selectionist and neutralist explanations, as selectionist explanations might also require conservation of hydrophobicity. However, it has the potential of falsifying a neutralist hypothesis.

Knowing whether more of the substitutions do this than expected is not, however, trivial. The code is arranged such that point mutations tend to conserve hydrophobicity (for quantification see Haig and Hurst, 1991). A bias to conservation is therefore expected from the null model that all non-synonymous mutations are equally likely to be fixed, regardless of hydrophobicity. Given, too, an ambiguity regarding transition/transversion rates (and hence the expected rate of different non-synonymous changes), predicting the null expectation for the degree of conservation is hard to do unambiguously. However, here we perform a simple, albeit rough, alternative test. It is known that mutations at the first site in a codon tend to conserve hydrophobicity where those at the second site do not (Haig and Hurst, 1991). Assuming that there is no reason to expect more mutations at the first rather than the second site, the neutralist model would be falsified by not finding an excess of mutations at the first rather than at the second site.

We have done this for all the non-synonymous substitutions in each signal peptide. We find that of 275 mutations, 159 affect the first site, while 116 affect the second, a significantly greater excess ($P<0.01$), consistent with expectations.

3.6. Rates of protein evolution in signal peptide and mature peptide are correlated

An earlier study of signal sequences (Smith and Hurst, 1998) indicated that the rate of evolution of the entire peptide may well be correlated with the rate of evolution of the signal peptide. In our dataset as well the K_A/K_S of the signal peptide strongly co-varies with the $K_A/K_S$of the entire peptide (r^2 ranked data=0.122, rank correlation $P=0.002$).

However, in this analysis [and the previous one (Smith and Hurst, 1998)] the signal sequence is included within the entire peptide, so introducing a non-independence. If we analyse the rates of evolution of signal peptides and compare them with those of the mature peptides, this non-independence is removed. We now fail to find a strong correlation between K_A/K_S of the mature and signal peptides, although there might be a tendency (ranked data $r^2=0.03$, $P=0.12$) (Table 1). Similarly, we find that the K_S values do not correlate ($P=0.19$). However, we find a strong positive correlation between the absolute rate of evolution of signal sequences (K_A) and that of mature peptides (regression of ranks: $P=0.001$).

3.7. Signal sequences of immune genes are fast evolving

Previous analyses have indicated that throughout their sequence, immune system genes are fast evolving (Hurst and Smith, 1999). Is the same true of their signal peptides? From a neutralist perspective it is hard to see why selection acting on the mature peptide should affect substitution rates in the signal peptide.

90

Comparing signal peptides of immune system genes ($N=14$) with non-immune genes, we found that the former tended to be faster evolving (assayed by K_A/K_S: Mann–Whitney U test, $P=0.03$; see also Appendix C). The immune genes' mature peptides are also faster evolving ($P<0.001$). Given that high rates of evolution through the rest of the sequence are most likely a result of antagonistic co-evolution, this finding is consistent with some high rates of evolution in signal peptides being associated (directly or indirectly) with the same. Analysis of intra-populational variation would be helpful to clarify the issue.

3.8. *Igf2r's signal peptide evolves at a rate comparable with that of many immune genes*

To determine whether the signal peptide of Igf2r evolves at a faster rate than other signal peptides, the signal peptide K_A/K_Sv alues were ranked. The K_A/K_S of *Igf2r*'s signal peptide is ranked 67 out of 76 (higher ranks being faster evolving). In large part this is because the K_S of the signal peptide is low: taking the K_A alone, it was ranked 50 out of all the 76 signal peptides. Neither statistic suggests that it is an outlier as previously indicated on the basis of a sample size an order of magnitude smaller.

We can also ask, given the rate of evolution of the mature peptide, does *Igf2r* have an unusually fast rate of evolution? In order to take this covariance into consideration, the difference in rank of the signal peptide K_A/K_S and the mature peptide was examined. *Igf2r* was found to have a positive difference in rank, but 11 genes had a higher difference, i.e. a higher K_A/K_S given the K_A/K_S of the mature peptide. This shows that, for the local K_A/K_S, *Igf2r*'s signal sequence is not an outlier. Likewise, its K_A, controlling for the K_A of the mature peptide is not unusual (25 have a greater K_A in the signal peptide given the K_A of the mature peptide).

Could the size of signal peptides be affecting this result? There is a much higher variation in signal peptide K_A/K_S (S.E.M. = 0.114, omitting three with a ratio of infinity) compared with entire peptide K_A/K_S (S.E.M. = 0.023). This could have been due to signal peptides being small and hence providing misleading estimates. If *Igf2r* has an unusually sized signal peptide, this might in part explain the findings. However we cannot substantiate this hypothesis.

By splitting the data set into two sets, one with large signal peptides and one with small, we found that there was no appreciable difference in K_A/K_S in the two sets (using two-tailed Mann–Whitney U test on signal peptide K_A/K_S, $P>0.05$). It was also shown that there was no difference in the variation between the two sets of data. Taking the squares of the residuals from the regression line (which was flat), we find the large and the small set are no different (two-tailed Mann–Whitney, $P>0.5$). These results all indicate that although the signal peptides are small, there is no trend with respect to their size within the group of signal peptides. Size effects are therefore unlikely to explain the K_A/K_S of *Igf2r* given the local rate of evolution. This result is further strengthened by noting that *Igf2r* itself has a signal peptide size of 96 bp, which is very close to the mean of the signal peptides in this data set (mean = 94.6).

All these results suggest that *Igf2r* signal peptide is probably not an outlier in our sample. However, as established, immune genes tend to have fast evolving signal sequences. Perhaps importantly, then, only four out of 14 immune genes have a higher K_A/K_S ratio in their signal peptides than *Igf2r* (Appendix C). This suggests that the rate of evolution of Igf2r's signal peptide may be, as originally claimed, unusually high for a non-immune gene.

That something unusual is going on is further supported by the finding that of six non-synonymous changes in the signal peptide, five occur at the second site, and do not conserve hydrophobicity. Four of the five reverse the hydrophobicity, as measured on the White interface scale (http://blanco.biomol.uci.edu/hydrophobicity_scales.html), the other causes a proportionally large change (the five second site changes are, in order 5′→3′: Q↔L, L↔P, R↔P, P↔L and L↔V). The number of non-synonymous changes at the first and second site is significantly different to that found in the other 75 genes taken in total (G-test of independence with Williams Correction = 4.06, $P<0.05$).

4. Discussion

This analysis set out to answer the following four questions:

1. Do signal peptides have rates of evolution expected of sequences that are perfectly neutral?
2. Does the rate of evolution of the signal peptide correlate with that of the mature peptide, possibly indicating a non-neutral force on peptide evolution?
3. Do all signal peptides evolve at the same rate?
4. Can we substantiate the claim that *Igf2r* has an especially high rate of evolution in its signal peptide?

As regards the first issue, it appears that signal peptides do evolve faster than the mature peptide, although by how much is dependent upon the measure. If we use paired samples, then they evolve on average between five and six times faster. However, if instead we take an average for all signal peptides and compare that with an average for all mature peptides, then they appear to evolve about twice as fast. Either way, a significant fraction of non-synonymous mutations must be under stabilizing selection. For the six lacking the usual hydrophobic core and that were on average significantly better conserved, the fraction must be much

higher. The functioning of the signal peptides of these unusual proteins is worthy of further investigation. Removal of the six signal peptides without the usual hydrophobic core does not affect the conclusion that signal sequences have many non-synonymous mutations under stabilizing selection (see Table 2).

The results above suggest that there could be different classes of signal peptides with different rates of evolution. However, although mitochondrial signal peptides are generally longer than non-mitochondrial ones, we failed to detect significant difference in the substitution rates of the two groups.

But what of the substitutions that we see, are these neutral? We could not falsify the hypothesis that the majority of non-synonymous mutations conserve hydrophobicity. We cannot therefore falsify the hypothesis that the substitutions are neutral. But neither does this permit us to falsify the hypothesis that they are under selection.

That neutral evolution may not be the only process going on in signal sequences is suggested by the fact that the absolute rate of protein evolution is correlated in mature and signal peptides. The best interpretation of the data that we can imagine is that there is some form of compensatory evolution going on: a change to the amino acids in the signal sequence might select for a change in the mature peptide or vice versa. While the activity of the mature peptide is independent of the signal peptide after the signal has been cleaved, prior to cleavage there may be selection on, for example, secondary or tertiary structure. There may, for example, be changes in the signal peptide that affect the activity of the mature peptide and/or vice versa. These would have to act prior to delivery of the mature peptide. Alternatively, the correlation between K_A for the mature and signal peptides might indicate genomic regional variation in the strength of the stabilizing selection.

We also find that signal peptides of immune system genes have unusually high rates of evolution. This is consistent with the hypothesis that some of the substitutions are driven by (or associated with) antagonistic co-evolution. It has been previously shown that coding regions of immune system genes tend to have high K_A/K_S ratios (Kuma et al., 1995; Hurst and Smith, 1999), a result that we can confirm. This can be accounted for by arguing that at least some part of the genes are under strong directional selection driven by host–parasite coevolution. It is perhaps surprising that the signal peptides of immune specific genes also evolve at an unusually high rate. This might however indicate, as before, that some of the frequent adaptive changes in the mature peptide regions cause slight disruptions in the secondary or the tertiary structure, that might select for compensatory changes in the signal sequences. Alternatively, one might speculate that the optimal cellular location (e.g., cytoplasm or membrane) of immune genes has been changed regularly as a response to new parasites.

Finally, we examined a particular selectionist hypothesis for the evolution of the signal sequence of *Igf2r*. Generally, were one to find rapid evolution (i.e. a high K_A/K_S ratio) of imprinted genes, it would provide reasonable support (McVean and Hurst, 1997) for the conflict theory for the evolution of imprinting, given that so many conflicts, for example maternal–foetal conflict (Hurst and McVean, 1998), do result in rapid evolution (but see also Haig, 1997). A previous study (McVean and Hurst, 1997) revealed that seven imprinted genes are not especially fast evolving. Further analysis by Smith and Hurst (1999) of 15 imprinted genes supported this broad conclusion, while noting that *Mash2* did have a rate of evolution comparable with immune system genes.

An earlier analysis (Smith and Hurst, 1998) indicated that Igf2r's signal peptide was an outlier, given the rate of evolution of the complete gene. We could find no evidence to indicate that this was an outlier in this larger data set, although its K_A/K_S was in the top 15% or so. However, for a non-immune gene it does appear to be fast evolving. Igf2r appears to have a signal peptide whose rate of evolution is higher than the majority of immune genes and comparable to that of the faster evolving ones. Given, too, that in the human–cow comparison the signal sequence also shows fast evolution (eliminating statistical artifact as an explanation), this result suggests that the rate of evolution of Igf2r's signal sequence might need special explanation. Examination of intra-population variation should help establish whether selection is acting on this sequence.

5. Conclusion

In summary, then, despite the fact that many random sequences function as signal peptides, we can certainly rule out the notion that signal peptides are paradigms of neutral evolution. Perhaps this is not surprising in retrospect, given that they are functional. Those putative signal peptides lacking the hydrophobic core evolve slowly at rates comparable to mature peptides. In part, this may be more a case of mistaken identity and an artifact of mis-annotation in GenBank. The remainder may be more nearly neutrally evolving than most sequences, but the unexpected correlation between the rate of protein evolution in the mature and the signal peptide suggests the unexpected possibility of compensatory evolution, suggesting that some of the non-synonymous substitutions could be the result of selection. This is supported by the finding that immune genes have high rates of evolution in their signal peptides and by the finding that Igf2r also has a fast evolving signal peptide for a non-immune gene.

92

Appendix A: The 76 mouse–rat orthologues and their substitution rates

Gene name (mouse)	Mouse Accession No.	Rat Accession No.	Mature peptide			Signal peptide			Signal peptide size	Rat cds size
			K_A/K_S	K_S	K_A	K_A/K_S	K_S	K_A		
Sterol carrier protein 2, liver[a]	M91458	M62763	0.068	0.291	0.020	0.45	0.057	0.026	60	432
Acetyl coenzyme A dehydrogenase, medium chain[a]	U07159	J02791	0.042	0.349	0.015	0.113	0.085	0.010	75	1266
Inhibin alpha[a]	X69618	M36453	0.103	0.228	0.024	0.17	0.166	0.028	699	1101
Inhibin beta-A[a]	X69619	M37482	0	0.112	0.000	0.27	0.111	0.030	924	1275
Coagulation factor III[a]	M26071	U07619	0.502	0.201	0.101	0.309	0.317	0.010	81	888
Ephrin B1[a]	U12983	U07560	0.061	0.138	0.008	0	0	0	23	1038
Small inducible cytokine A11[b]	U26426	U96637	0.189	0.075	0.014	1.596	0.041	0.065	69	294
Small inducible cytokine B subfamily, member 5[b]	u27267	u90448	0.694	0.474	0.683	0.719	0.163	0.117	120	393
Oxytocin[b]	m88355	m67442	0.076	0.194	0.015	0	0.131	0	57	378
Interleukin 4 receptor, alpha	m29854	x69903	0.708	0.179	0.127	0.544	0.152	0.083	75	2412
Low density lipoprotein receptor	x64414	x13722	0.263	0.238	0.063	0.374	0.124	0.046	63	2640
Tumour necrosis factor receptor superfamily, 1a	M60468	M63122	0.49	0.197	0.097	0.302	0.064	0.019	87	1386
Calreticulin	x14926	x53363	0.067	0.109	0.007	0	0.059	0	51	1251
Glutamate dehydrogenase	x57024	x14223	0.022	0.212	0.005	0.129	0.159	0.021	159	1677
Cathepsin E	x97399	D38104	0.241	0.152	0.036	0.878	0.295	0.259	57	1098
Insulin-like growth factor binding protein 5	x81583	m62781	0.015	0.105	0.002	0.556	0.051	0.028	57	816
Thyroid stimulating hormone receptor	u02602	m34842	0.13	0.218	0.028	0.761	0.185	0.141	63	2295
Gamma-aminobutyric acid receptor, subunit gamma 2[b]	m86572	l08497	0.019	0.219	0.004	0.504	0.048	0.024	114	1401
Gamma-aminobutyric acid receptor, subunit alpha 1[b]	m86566	l08490	0	0.148	0.000	0	0.165	0	141	1368
Myelin/oligodendrocyte glycoprotein (MOG)	u64572	m99485	0.108	0.203	0.022	0.442	0.268	0.118	81	738
Activin A receptor type II-like kinase 1	l48015	l36088	0.085	0.154	0.013	0.593	0.205	0.122	63	1515
Activin A receptor type II-like kinase	L15436	L19341	0.056	0.189	0.011	∞	0	0.63	45	1530
KGF-7[b]	u58503	x56551	0.123	0.124	0.015	0.264	0.168	0.044	75	585
Insulin like growth factor 2 receptor[b]	u04710	u59809	0.176	0.188	0.033	1.424	0.069	0.098	96	7449
Decay accelerating factor 1[b]	l41366	af039583	0.813	0.229	0.186	0.663	0.265	0.176	102	1200
Beta-glucuronidase structural	J02836	m13962	0.253	0.237	0.060	0.683	0.251	0.171	66	1947
Endothelin-1	D43775	m64711	0.214	0.262	0.056	0.504	0.062	0.031	51	609
Glycoprotein hormones, alpha subunit	M22992	j00757	0.077	0.227	0.017	1.535	0.044	0.067	69	363
Carboxyl ester lipase[b]	u33169	m69157	0.228	0.195	0.044	0	0.136	0	60	1839
Surfactant associated protein D	l40156	m81231	0.231	0.188	0.044	0.222	0.235	0.052	57	1125
5′ nucleotidase	L12059	J05214	0.149	0.185	0.028	1.094	0.118	0.129	84	1731
Secretory granule neuroendocrine protein 1, 7B2 protein	X15830	M63901	0.018	0.114	0.002	0.673	0.148	0.010	72	633
Insulin receptor	J05149	M29014	0.024	0.19	0.005	0.46	0.98	0.451	78	4152
Lysosomal membrane glycoprotein 1	M25244	M34959	0.465	0.203	0.094	1.249	0.217	0.271	63	1224
Luteinizing hormone receptor	M81310	M26199	0.148	0.1901	0.028	0.685	0.131	0.090	78	2103
Leukemia inhibitory factor	X12810	M32748	0.209	0.265	0.055	0.194	0.127	0.025	66	609
Mannose binding lectin, serum (C)	D11440	M14103	0.645	0.202	0.131	0.117	0.817	0.096	54	735
Myelin-associated glycoprotein	M31811	M16800	0.096	0.165	0.016	0.135	0.231	0.031	48	1881
Matrix gamma-carboxyglutamate (gla) protein	D00613	J03026	0.412	0.174	0.072	∞	0	0.03	57	312
Leptin	U18812	D45862	0.184	0.101	0.019	0	0.217	0	63	504
Secreted phosphoprotein 1	X16151	M14656	0.397	0.214	0.085	0.74	0.163	0.121	66	954
Transthyretin	D00073	K03252	0.155	0.261	0.040	0.249	0.217	0.054	60	444
Pancreatitis-associated protein	D13509	M55149	0.166	0.3	0.050	0.177	0.632	0.112	78	528
Parathyroid hormone receptor	X78936	M77184	0.049	0.132	0.006	0.358	0.071	0.025	63	1776
Uteroglobin	L04503	J05536	0.624	0.111	0.069	0.338	0.063	0.021	57	291
Pancreatic polypeptide	M18208	M13588	0.195	0.493	0.096	0.607	0.087	0.053	87	297
Prolactin receptor	L13593	M57668	0.319	0.148	0.047	0.237	0.297	0.07	114	1833
Selectin, platelet	M87861	L23088	0.228	0.227	0.052	0.326	0.271	0.088	123	2307
Rat regenerating islet-derived, mouse homologue 1	D14010	M62930	0.439	0.157	0.069	0.281	0.455	0.128	63	498
Insulin-like growth factor binding protein 6	X81584	M69055	0.224	0.13	0.029	0.132	0.314	0.041	75	681
CD1d1 antigen	M63695	D26439	0.397	0.221	0.086	1.801	0.057	0.103	54	1011
Cytochrome C oxidase, subunit Vb	X53157	D10952	0.19	0.134	0.025	0.632	0.169	0.107	93	390
Secreted acidic cysteine rich glycoprotein	X04017	D28875	0.11	0.136	0.015	0	0	0	51	906
Mast cell protease 7	L00653	D38455	0.302	0.643	0.201	0.781	0.417	0.326	57	825
Granzyme B	X04072	M34097	0.542	0.204	0.111	1.232	0.134	0.165	60	747
Kallikrein-3, plasma	M58588	M58590	0.275	0.189	0.052	0.2	0.254	0.051	57	1917
Receptor tyrosine kinase	U18933	D37880	0.119	0.125	0.015	0.119	0.13	0.015	93	2643
Thrombopoietin	L34169	D32207	0.621	0.131	0.081	0.932	0.185	0.172	63	981
CD3 antigen, zeta polypeptide	J04967	D13555	0.217	0.118	0.026	0.214	0.431	0.092	63	495
TGF-alpha	M92420	M31076	0.03	0.112	0.003	1.038	0.025	0.026	114	480
Acid phosphatase 5, tartrate resistant	M99054	M76110	0.098	0.25	0.025	0.689	0.243	0.167	63	984
UDP-glucuronosyltransferase 1 family, member 1	U09930	J02612	0.191	0.184	0.035	0.058	0.232	0.013	210	1590
Mouse vasopressin-neurophysin II	M88354	M25646	0.103	0.176	0.018	0.314	0.168	0.053	57	495
Lymphocyte antigen 84	Y07519	U04319	0.587	0.176	0.103	0.69	0.194	0.134	78	1011
Small inducible cytokine A5	X70675	U06436	0	0.161	0.000	∞	0	0.109	66	279
Follistatin-like polypeptide	M91380	U06864	0.122	0.16	0.020	0.271	0.09	0.024	54	921
Carbonic anhydrase 5, mitochondrial	X51971	U12268	0.345	0.216	0.0750	2.085	0.072	0.150	102	915
Biglycan	L20276	U17834	0.021	0.108	0.002	0.683	0.046	0.031	57	1110
Immunglobulin CTLA-4	X05719	U37121	0.233	0.157	0.037	1.79	0.054	0.010	111	672
Acetyl coenzyme A dehydrogenase, short chain	L11163	J05030	0.039	0.261	0.010	0.261	0.422	0.110	78	1245
Orosomucoid 1	M27008	J00696	0.677	0.24	0.162	0.443	0.279	0.124	54	618
Islet amyloid polypeptide	M25389	J04544	0.11	0.243	0.027	0.612	0.159	0.097	111	282
Apolipoprotein A-IV	M64249	M00002	0.449	0.235	0.106	1.2	0.068	0.082	60	1176
Calcium binding protein, intestinal	J05186	M86870	0.133	0.168	0.022	0	0	0	72	1932
Casein kappa	M10114	K02598	1.5	0.102	0.153	0	0.275	0	63	537
Matrix metalloproteinase 7	L36244	L24374	0.284	0.223	0.063	7.771	0.026	0.202	60	804

[a] The first six entries are those genes with signal peptides with unusual hydrophobicity plots.

[b] The mouse signal peptide was used for the analysis, due to the lack of annotated signal peptide region of the rat orthologue.

Appendix B: Mitochondrial signal peptides

Gene name (mouse)	Mouse Accession No.	Rat Accession No.	K_A/K_S mature peptide	K_S mature peptide	K_A/K_S signal peptide	K_S signal peptide	Signal peptide size
Carbonic anhydrase	X51971	U12268	0.345	0.216	2.085	0.072	102
Ornithine carbamoyltransferase	M17030	K00001	0.079	0.108	1.49	0.06	96
ATP synthase alpha subunit	L01062	J05266	0.026	0.306	0.809	0.056	99
ATP synthase coupling factor 6	U77128	M73030	1.469	0.04	0.58	0.116	96
Malate dehydrogenase	M16229	X04240	0.049	0.206	0.224	0.095	72
Aspartate aminotransferase isoenzyme	J02622	J02622	0.029	0.183	0.203	0.177	87
ATP synthase subunit c	L19737	D13123	0.062	0.11	0.163	0.141	183
FAD-linked glycerol-3-phosphate dehydrogenase (Gdm1)	U60987	U08027	0.097	0.249	0.264	0.097	126

Appendix C: The 10 fastest evolving (highest K_A/K_S) signal peptides

Ranked K_A/K_S of signal peptide regions	Name	Expression pattern/effect of mutation	Classification
67	Igf2r	Embryo, placenta, nervous system, etc.	Biochemical: receptor
68	Glycoprotein hormones, alpha subunit	Produced in both gonadotrophs and thyrotrophs/ endocrine defects, growth defects, obesity	Physiological: glycoprotein hormone
69	Small inducible cytokine A11	Immune system	
70	Immunglobulin CTLA-4	Immune system (immunoglobulin superfamiliy)	Glycoprotein
71	CD1d1 antigen	Immune system (CD1 antigen)	Surface glycoprotein
72	Carbonic anhydrase 5, mitochondrial	Housekeeping gene	Mitochondrial biochemical enzyme
73	Matrix metalloproteinase 7	Thymus, speen, liver, placenta, uterus mammalian gland	Biochemical: enzyme
74	Activin A receptor type II-like kinase 2	Embryo (growth factor receptor?)	Biochemical: receptor
75	Matrix gamma-carboxyglutamate (gla) protein	Osteoblasts during embryogenesis	Biochemical: emzyme
76	Small inducible cytokine A5	Immune system	

Acknowledgement

We wish to thank the ESF's TBA program for providing resources for C.P. L.D.H. is funded by the Royal Society.

References

Aguade, M., Miyashita, N., Langley, C.H., 1992. Polymorphism and divergence in the mst26a male accessory-gland gene region in *Drosophila*. Genetics 132, 755–770.

Dickerson, R.E., 1971. The structure of cytochrome c and the rates of molecular evolution. J. Mol. Evol. 1, 26–45.

Duret, L., Mouchiroud, D., Gouy, M., 1994. HOVERGEN — a database of homologous vertebrate genes. Nucleic Acid. Res. 22, 2360–2365.

Freeland, S.J., Hurst, L.D., 1998. Load minimization of the genetic code: history does not explain the pattern. Proc. R. Soc. Lond. B 265, 2111–2119.

Gillham, N.W., 1995. Organelle Genes and Genomes. Oxford University Press, Oxford.

Haig, D., 1997. Parental antagonism, relatedness asymmetries and genomic imprinting. Proc. R. Soc. Lond. B 264, 1657–1662.

Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization in the genetic code. J. Mol. Evol. 33, 412–417.

Hughes, A.L., 1991. Circumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions. Genetics 127, 345–353.

Hughes, A.L., 1992. Positive selection and interallelic recombination at the merozoite surface antigen-1 (msa-1) locus of plasmodium-falciparum. Mol. Biol. Evol. 9, 381–393.

Hughes, A.L., Ota, T., Nei, M., 1990. Positive Darwinian selection promotes charge profile diversity in the antigen binding cleft of class I MHC molecules. Mol. Biol. Evol. 7, 515–524.

Hurst, L.D., McVean, G.T., 1998. Do we understand the evolution of genomic imprinting? Curr. Opin. Genet. Dev. 8, 701–708.

Hurst, L.D., Smith, N.G.C., 1999. Do essential genes evolve slowly? Curr. Biol. 9, 747–750.

Izard, J.W., Kendall, D.A., 1994. Signal peptides — exquisitely designed transport promoters. Mol. Microbiol. 13, 765–773.

Kaiser, C.A., Preuss, D., Grisafi, P., Botstein, D., 1987. Many random sequences functionally replace the secretion signal sequence of yeast invertase. Science 235, 312–317.

Kuma, K., Iwabe, N., Miyata, T., 1995. Functional constraints against variations on molecules from the tissue-level — slowly evolving

brain-specific genes demonstrated by protein-kinase and immunoglobulin supergene families. Mol Biol Evol. 12, 123–130.
Ladunga, I., 1999. PHYSEAN: PHYsical SEquence ANalysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. Bioinformatics 15, 1028–1038.
Li, W.-H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.
Li, W.-H., Gojobori, T., Nei, M., 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292, 237–239.
Li, W.-H., Wu, C.-I., Luo, C.-C., 1985. A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2, 150–174.
Makalowski, W., Boguski, M.S., 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA 95, 9407–9412.
McVean, G.T., Hurst, L.D., 1997. Molecular evolution of imprinted genes: no evidence for antagonistic coevolution. Proc. R. Soc. Lond. B 264, 739–746.
Moore, T., Haig, D., 1991. Genomic imprinting in mammalian development: a parental tug-of-war. Trends Genet. 7, 45–49.
Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1–6.
Pamilo, P., Bianchi, N.O., 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol. Biol. Evol. 10, 271–281.
Smith, N.G.C., Hurst, L.D., 1998. Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. Genetics 150, 823–833.
Smith, N.G.C., Hurst, L.D., 1999. The causes of synonymous rate variation in the rodent genome: Can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152, 661–673.
Tan, Y., Riley, M.A., 1997. Positive selection and recombination: major molecular mechanisms in colicin diversification. Trends Ecol. Evol. 12, 348–351.
Tsaur, S.C., Wu, C.I., 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. Mol. Biol. Evol. 14, 544–549.
Wolfe, K.H., Sharp, P.M., 1993. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J. Mol. Evol. 37, 441–456.

95

Chapter 8: Natural selection promotes the conservation of linkage of co-expressed genes

Laurence D. Hurst, Elizabeth J. B. Williams and Csaba Pal (2002)

Trends in Genetics, submitted

Whilst there is increasing evidence that eukaryotic gene order is not always random[1-5], there is no evidence that putatively favourable gene arrangements are preserved by selection more than expected by chance. In yeast (*Saccharomyces cerevisiae)*, for example, co-expressed genes tend to be linked[2,6,7] but it is unknown if such gene pairs tend to remain linked more often than expected under null neutral expectations. We show using gene pairs in the yeast-*Candida* comparison, that highly co-expressed gene pairs are conserved as pairs at about twice the average rate. Whilst, as expected from a null neutral model, genes that tend to be closer together are retained more commonly and co-expressed genes tend to be in close physical proximity, this closeness only accounts for a small proportion of the enhanced degree of conservation of co-expressed gene pairs. These results demonstrate that purely neutralist models of gene order evolution are not realistic.

Much current data suggests that the randomly arranged beans on a string model of eukaryotic genomes is not adequate. Not only are certain sorts of genes especially prevalent on the X chromosome[8,9], but in humans[5], flies[4], yeast[2] and worm[3] genes of similar expression profile tend to be clustered. In striking contrast, there is very little evidence to suggest that, with obvious exceptions such as hox clusters[10], any putatively adaptive clusters remain conserved more often than expected of any random set of genes. Based on a limited sample it has, however, been suggested that co-expressed genes in yeast might be conserved at a higher rate than expected[11], although a broad scale analysis failed that show that gene orientation (a putative covariate of co-expression), was biased in conserved gene pairs[12].

To address this issue we assembled a dataset of gene pairs (i.e. nearest neighbours) from *Saccharomyces Cerevisae,* for which we could define the ortholog for both in *Candida* (1850 pairs). Orthology was determined using reciprocal best hits in Blast analysis, as previously described[13]. Chromosomal location in *S. cerevisae* was derived from accession numbers NC_001133-48. Protein sequence and location data for *C. albicans* was obtained from the Stanford DNA Sequencing and Technology Center website at http://www-sequence.stanford.edu/group/Candida; contig version 6.

Of the 1850 *Saccharomyces* gene pairs with *Candida* orthologs, we eliminated those that were pairs of tandem duplicates (as defined by pairwise blast score $E<10^{-2}$), those that were overlapping or with no space between the genes and those for which we could not define the extent of co-expression between neighbouring genes. This left a total of 1817 genes. The data set comprised 166 gene pairs in *Saccharomyces* that remain as nearest neighbours in *Candida*. These we consider to be the gene pairs with conserved linkage. The overall proportion conserved (9%) is low, but this is

more a measure of the long time since common ancestry (about 200 million years) than an indication of the presence/absence of selection. Indeed comparison can be made with the evolution of codon usage bias: in *Saccharomyces* highly expressed genes show strong codon usage bias indicative of selection acting on "silent" point mutations, but in comparisons of these genes with orthologs in *Candida* the silent site substitution rate is very high and so close to saturation as to be uninterpretable.

To establish if co-expression is important for retention of linkage we need to define the extent of co-expression. We took the expression profiles of the *Saccharomyces* genes from the microarray data compiled by the Eisen lab (http://rana.lbl.gov/EisenData.htm) and, using normalised data, for each linked pair calculated the Pearson correlation coefficient (r) between the two genes, a measure of their degree of co-expression. If co-expression were to be important in the retention of a gene pair, then we should expect that as the degree of co-expression goes up, so too the probability of conservation of linkage should increase. However, we have no reason to suppose that this need be a gradual effect. For most gene pairs the r values simply represent random noise: a small positive value for r should not be taken as evidence of more co-ordinated expression than an equally small negative value. Only when the r value is especially high do we suspect some functionally significantly co-ordination in the regulation of the two genes. Therefore, to provide an indication as to whether co-expression is of importance, we performed a sliding window analysis of gene pairs organised by the ranked r value, calculating mean r, the proportion conserved and mean intergene spacer. As can be seen in figure 1, at high values of mean r (highly co-expressed genes), the proportion conserved does indeed exceed greatly null expectations. This provides the first whole genome analysis to indicate that co-expressed genes are conserved more then expected by chance. As expected

then, the genes pairs that are conserved have higher r values than those not conserved (Mann Whitney U test P=0.01).

There is, nonetheless, a difficulty with the interpretation of the above result. Examination of figure 1 also indicates that as mean r tends to increase, so too, mean intergenic distance tends to decrease. The excess conservation of co-expressed gene pairs might then trivially be explained as a consequence of a null neutral evolution of gene order. The simplest explanation for the conservation of linkage is that gene order re-arrangements (e.g. inversions) occur at random locations, that they are tolerated only if they disrupt intergene spacer and that all such tolerated re-arrangements are without selective consequences. The tolerated ones then may spread by drift (i.e. neutral evolution). Gene pairs with small intergene spacer should then be expected to be conserved as nearest neighbours more often. Indeed, as predicted, conserved gene pairs are closer together than non-conserved ones (Mean intergene spacer unconserved pairs = 510.8 bp, ±16.3; Mean intergene spacer of conserved pairs = 333.0 ±17.3: Mann Whitney U test, $P < 0.0001$).

Further, a sliding window analysis reveals a general tendency for the probability of gene pair conservation to decline as intergene spacer increases (Figure 2). A conservative set of non-co-expressed gene pairs (N=1124), also shows the same decline in proportion of conserved gene pairs as intergene spacer increases (Figure 2). This indicates that the decline in probability of conservation with intergene spacer size is not simply owing to co-expressed genes both being more highly conserved and having to have small intergene spacer, possibly to ensure optimal co-regulation. Tests for conservation of linkage should then control for intergene distance.

Does the physical proximity of co-expressed genes account in full for their preferential conservation? We need to start by defining a set of genes as being co-

expressed. To some extent this is arbitrary, but by examination of figure 1, we define the 250 genes with the highest *r* values as being co-expressed. This corresponds approximately to those samples in figure 1 with mean *r* greater than 0.52 and represents the approximate position at which the proportion conserved hits the apparent asymptote. This is probably a conservative definition, but as it is our intention to ask whether we can detect a signal of selection, utilizing the uppermost 14% of genes (in terms of co-expression) is adequate. In this set 42 are conserved (16.8%) as opposed to 124 of the remaining 1567 (8%) (G test of independence: G_{adj}=13.91, P<<0.001, P< 0.00005 from 50,000 randomizations). These co-expressed genes have significantly smaller intergene spacer (Mean intergene spacer co-expressed pairs = 446.5 bp, ±36.2; Mean intergene spacer of remaining pairs = 502.2 bp ±16.4: Mann Whitney U test; *P*=0.03).

To examine whether this higher degree of conservation was owing to the reduced intergene spacer size, we split the data into non-overlapping groupings of approximately equal intergene spacer (1-200bp, 201-400 bp etc) and for each sub-group we determined the number of co-expressed genes that are conserved. In each subgroup, the expected number is the number of co-expressed gene pairs (conserved or not conserved) multiplied by the proportion of genes conserved in the data set as a whole within the same subgroup. In each subgroup, the numbers of co-expressed gene pairs conserved is higher than null expectation and is, overall, very significantly higher than expected (χ^2=17.0, *P*=0.004) (Figure 3). It appears then that the most highly co-expressed genes are indeed conserved more commonly than expected by chance, even allowing for their physical proximity.

How important is selection compared to neutral evolution in the preservation of the most highly co-expressed class? If we ignore the effect of intergene spacer we

expect 166 (total conserved) /1817 (total genes) x 250 highly coexpressed genes, i.e. 22.84. We therefore would report an excess of 19.16 in the co-expressed class. When instead we estimate the excess found when we control for intergene spacer length, by counting up the expected number of conserved genes in each block (0-200bp etc) and comparing this to the observed (42), we find an excess of 18. Therefore the control for intergene spacer explains relatively little of the excess conservation of co-expressed genes (approximately 5%). This also reflects the fact that the mean intergene spacer is not greatly significantly lower in the co-expressed class, while the apparent enrichment of conservation of gene pairs is highly significant.

Prior evidence suggests that divergently oriented genes (← →) are especially likely to belong to a single regulatory unit[7]. Within the highly co-expressed group, the genes in divergent orientation are indeed more common than expected from their overall frequency (we expect 65 but observe 85). Contrary to prior suggestions[7], we find a dearth of gene pairs in which both genes are in the same orientation (87 observed ,115 expected). Overall, there is a significant difference in the proportion of types in different orientations in the highly co-expressed class compared with their frequencies in the data set as a whole (χ^2=13.88, ν=2, P<0.001).

Of the 42 pairs that are conserved within the highly co-expressed class, 19 (45%) are in the divergent orientation in yeast, approximately double their frequency within the data set as a whole (G test of independence, P<0.01), and higher than their frequency within the co-expressed class, although not significantly so (G test of independence, P>0.05). While the above results suggests that divergent orientation is important for co-regulation and for conservation of pairs, we do not find that the divergent genes retain their orientation at an especially high rate. Of the 19, 14 (74%) have the same orientation in *Candida*, which compares with 62% of conserved pairs

that have the same direction in both species (i.e. 103 out of the sample of 166). Nonetheless, our findings are consistent with the observation[14] that between *S. cerevisiae* and *Candida albicans*, divergently transcribed gene-pairs that are conserved in evolution have a higher probability of being co-regulated than divergently transcribed gene pairs that are disrupted in evolution.

We conclude that, consistent with the null neutral model gene pairs that have small intergene spacer are the most highly conserved. This result emphasises the need to control for the length of intergene spacer when testing hypotheses of gene order evolution. However, amongst the most highly co-expressed gene pairs a clear signal of selection is evident, with co-expressed genes being retained at about twice the expected rate. This can only in small part be explained as a consequence of reduced intergene spacer. Consequentially the null neutral model cannot be considered an adequate description of gene order conservation.

Acknowledgements

We should like to thank Martijn Huynen and an anonymous referee for comments on a prior version of the manuscript. CP is funded by a Royal Society/Nato visiting fellowship and LDH by B.B.S.R.C

References

1 Hurst, L.D. (1999) *Trends Ecol. Evol.* 14, 108-112

2 Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M. (2000) *Nature Genetics* 26, 183-6.

3 Blumenthal, T. et al. (2002) *Nature* 417, 851-4

4 Spellman, P.T. and Rubin, G.M. (2002) *J. Biol.* 1, 5

5 Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) *Nature Genetics* 31, 180-183

6 Cho, R.J. et al. (1998) *Mol Cell* 2, 65-73

7 Kruglyak, S. and Tang, H. (2000) *Trends Genet* 16, 109-11

8 Wang, P.J., McCarrey, J.R., Yang, F. and Page, D.C. (2001) *Nature Genetics* 27, 422-426

9 Saifi, G.M. and Chandra, H.S. (1999) *Proceedings of the Royal Society of London Series B-Biological Sciences* 266, 203-209

10 Zhang, J.Z. and Nei, M. (1996) *Genetics* 142, 295-303

11 Huynen, M.A., Snel, B. and Bork, P. (2001) *Trends Genet* 17, 304-6

12 Seoighe, C. et al. (2000) *Proc Natl Acad Sci U S A* 97, 14433-7

13 Pal, C., Papp, B. and Hurst, L.D. (2001) *Molecular Biology and Evolution* 18, 2323-2326

14 Huynen, M.A. and Snel, B. in *Frontiers in Computational Genomics* (Galperin, M.Y. and Koonin, E.V., eds), pp. (in press), Horizon Scientific Press

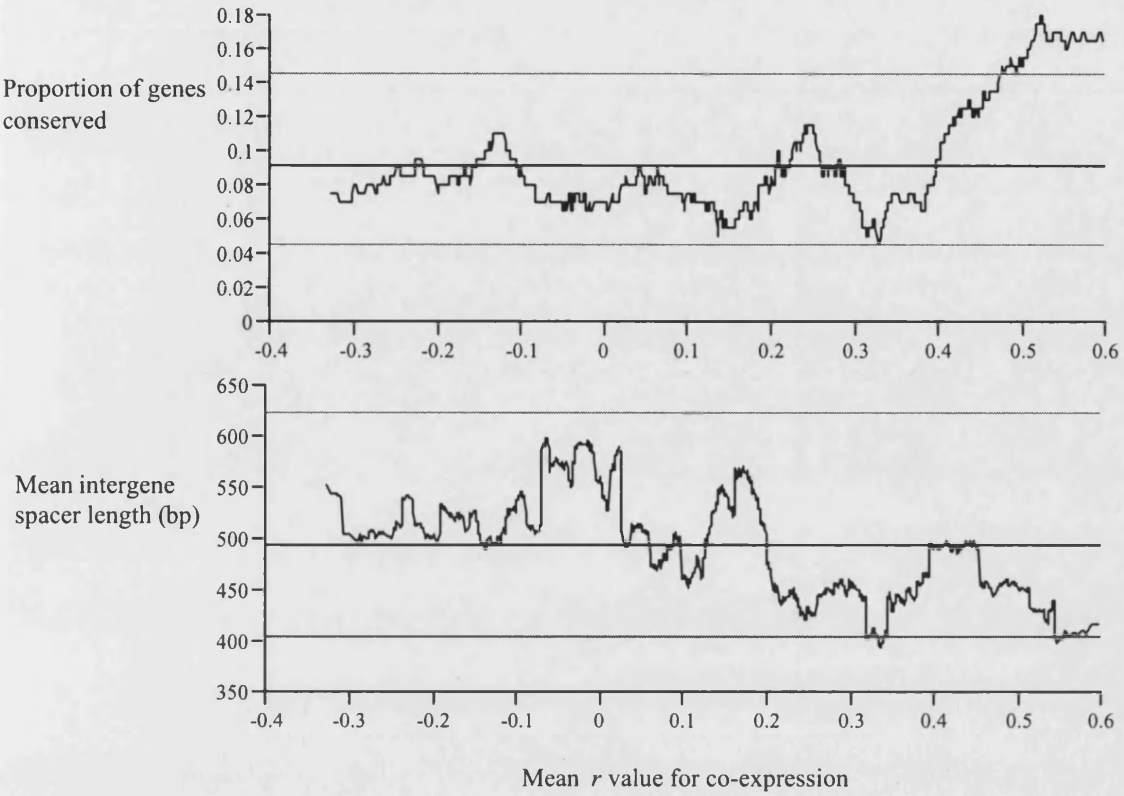


Figure 1. The proportion of genes conserved and mean intergene spacer length as a function of the mean degree of co-expression (*r* value) for groupings of 200 genes. All the 1817 genes [you mean gene pairs] were ranked by *r* value (rank of 1 = most highly co-expressed gene pair). We then examined the first 200 genes [again gene pairs] (ranks 1-200) and determined mean *r*, mean intergene spacer and the proportion conserved. The right most data on the two plots represents the data from this group. We then moved to the genes ranked 2 to 201, then 3 to 202 etc., and in each calculated the same parameters. In each plot the solid black line indicates the mean and the two grey lines indicate the 99% confidence interval determined by randomization. These were generated by random sampling of 200 of the 1817 genes without replacement, 10,000 times.

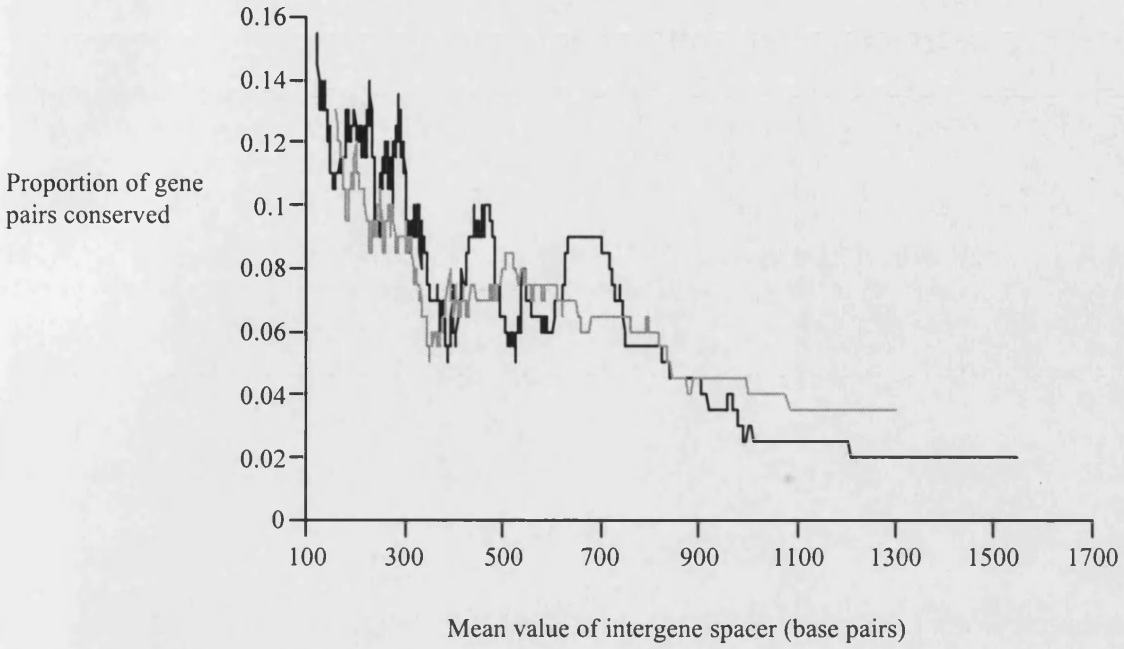


Figure 2 The proportion of gene pairs seen in yeast that are conserved as a pair in *Candida* as a function of the length of the intergene spacer between the genes in yeast. The black line indicates the data for the dataset as a whole. The grey line indicates the data for those that are not co-expressed (N=1124). The figure is a sliding window plot of the data ranked by intergene spacer size using 200 gene pair samples and a jump of one gene pair between windows, as in figure 1. The "non co-expressed" set was deduced from a *P* value for the significance of the observed *r* value between two genes derived by randomization. Those with $P>0.05$ were assumed to not be co-expressed. This provides a conservative sample of non-co-expressed genes as, owing to multiple sampling, there will be numerous gene pairs that show spurious significance at the 5% level, but are excluded from this data set.

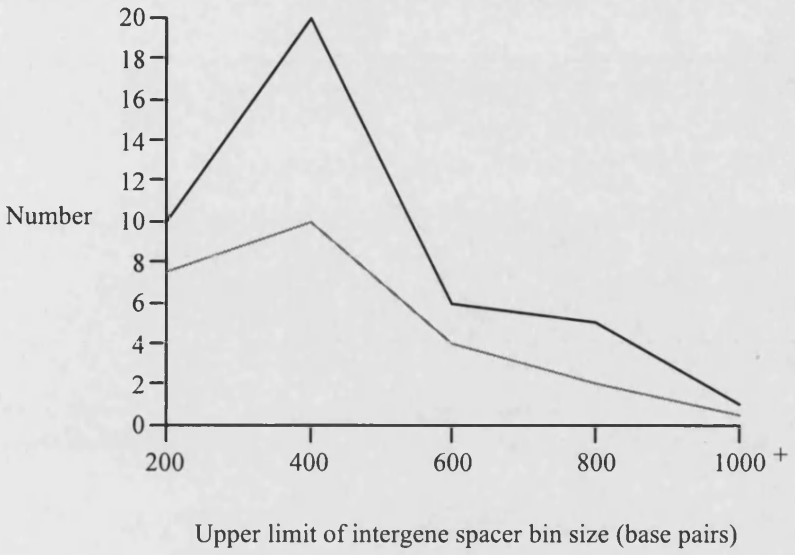


Figure 3. The number of conserved highly co-expressed gene pairs observed (black) and expected (grey) as a function of the intergene spacer size. The X axis figures indicate the upper limit to the subgroup size (i.e. 200 represents genes with intergene spacer less than or equal to 200, 400, indicate 201 to 400 etc.).

Chapter 9: Discussion

In the introduction I showed that there are two possible explanations for the existence of variation seen in populations, both in the visible, outwards attributes of species but also in the less easily seen genome sequence. The two explanations are that the variation is due either 1) to random mutational processes according to the neutralist explanation or 2) to selection. This has been a long running debate and as such many theories have been developed, proven, disproved and re-evaluated in support of either side.

I have attempted to address this debate by looking at the variation in rates of evolution and what factors predict the rate of evolution. I hope that some of what I have discovered contributes to the debate between the two opposing camps. The finding that the location of a gene is related to the rate of protein evolution, the rate of silent site evolution, its expression profile as well as the rate of chromosomal rearrangements seems to suggest that the null model of genes being randomly distributed around the genome cannot be sustained. Selective pressure must be used to explain some of the variation seen in genomes.

Looking at the conservation of gene order, I found that gene pairs were more likely to be conserved if 1) they were closer together and 2) they had similar expression profiles. This was evidence for both the neutral and the selection sides to the debate. However, the suggestion that genome organisation is under selection is also supported by my finding that linked genes in rodents

have similar expression profiles. A much larger study of human genes confirmed this finding (Lercher et al., 2002).

This suggests that gene clusters such as the HOX cluster and the mammalian major histocompatability cluster (MHC) may not be exceptions. It is possible that linked genes share transcriptional control regions, especially those in divergent orientations (Kruglyak and Tang, 2000). In addition, in *Caenorhabditis elegans* there are examples of operons (Blumenthal, 1998; Blumenthal et al., 2002). These were originally thought to be unique to eubacterial genomes. Are these examples the tip of the iceberg? Is there more to be discovered about the evolution of genome organisation, especially the evolution of gene expression control in eukaryotes? The vast amount of microarray, SAGE and other large scale expression data becoming available should hopefully allow us to investigate the relationship between genome organisation and gene expression in detail.

Examining the evolution of gene expression will probably require further understanding of the processes surrounding the control of gene expression. At the moment little is known about how control elements evolve and whether selection or neutralist forces predominate. Indeed, it may come down to a chicken and the egg scenario, did the transcription factor change first or did the control element? In addition, genome organisation may not be simply a one-dimensional array of genes along a chromosome, but may in fact be three dimensional in nature with interactions occurring between genes on different chromosomes. The evolution of chromatin organisation is a fascinating area which is only recently being looked at by evolutionary biologists.

However, the other patterns I looked at gave a much more complicated picture than those mentioned above. The expectation that the similarity of rates of protein evolution was due to linkage effects, such as the Hill-Robertson effect (Hill and Robertson, 1966) seems to be far from conclusive. In fact separating such effects from the finding that linked genes have similar expression patterns will require much more analysis. However, the major area of research in molecular evolution in the future, I would like to suggest, will most likely be explaining the variation in Ks especially in light of recent findings.

Let us begin our examination of the Ks debate by looking at my finding that there is large-scale local similarity in rates of silent site substitution (chapter 3). The discovery that there is large chromosomal variation in Ks was surprising. If, as was expected, most of the variation in mutation rates is due to time spent in the male germ line, then there should have been little variation in Ks between autosomes. The finding of large differences in Ks on different autosomes was not at all expected. Castresana, repeating my results, suggested that chromosomal heterogeneity in Ks could be due to GC% variation around the genome (Castresana, 2002). Indeed the lack of local similarity in Ks in the rodent genome appears to be a real effect and this too could be due to the lower variance in GC% seen in the rodent genome. The rodent genome has also been shown to have unusual patterns of substitution (Robinson et al., 1997). However, when we controlled for GC% (chapter 3), we still found local similarity in Ks. Therefore, variation in GC% cannot explain all the local similarity in rates of silent substitutions.

Adding to the confusion in the literature surrounding Ks there is growing evidence that estimates of Ks, even using maximum likelihood methods, are

biased by the nucleotide composition of the sequence they are analysing (Pesole et al., 1995). The resulting correlation of GC% and Ks varies considerably depending on the method used to estimate the silent site substitution rate (Smith and Hurst, 1998; Bielawski et al., 2000). In addition, we found that the repeatability of the rate of silent substitutions depends heavily on the method used to calculate Ks, and K4 shows no evidence of repeatability.

Using K4 as an estimate of the silent substitution rate solves many of the problems associated with estimating Ks. Using Tamura and Nei's model on four fold sites I get a positive correlation that shows only a small increases in K4 with GC (Human-mouse orthologues, N = 4365, r^2 = 4.9%, P < 0.001)(Tamura, 1992). In comparison Goldman and Yang's maximum likelihood method (Goldman and Yang, 1994) gave an accelerating power function relating Ks to GC when looking at over 4000 human – mouse orthologues (Castresana, 2002). Both methods cannot be right. It may be useful to closely examine the methodology used to calculate Ks and to see if there are any major differences between the methods, such as in the assumptions that they are based on.

Since estimating rates of silent substitutions is problematic, it becomes harder to determine what is causing the variation seen in Ks. This leads to a difficulty in determining whether base composition, or isochore structure, is due to selective or mutational pressures.

The situation is further complicated by recent evidence suggesting that this isochore structure in mammals is degrading (Duret et al., 2002). It was found that there is an excess of GC->AT substitutions in GC rich genes. Duret suggests that biased gene conversion is a probable cause for the preference of

GC substitutions over AT in regions of high GC due to analysis of polymorphism data. This may be related to the finding by Kumar & Subramanian (Kumar and Subramanian, 2002) that genes with different substitution patterns in different lineages tend to be GC rich. This suggests that the non-equilibrium nature of isochores is responsible for the difference in substitution patterns, however this has not yet been tested.

The possibility that the isochore structure in mammals is degrading is contrary to the assumption many models make that base composition is stationary, such as the rejection of mutation bias by Eyre-Walker (Eyre-Walker, 1999; Smith and Eyre-Walker, 2001). He claimed that SNP data failed to support mutation bias models of isochore evolution but he assumed that base composition was stationary (Smith and Eyre-Walker, 2001). The finding of non-stationarity adds an extra complication to the study of the evolution of nucleotide variation in genomes. Therefore, conclusions that assumed stationary base composition need to be re-evaluated. This includes some of those presented in this thesis. For example, my finding that there is a positive correlation between Ks and GC may simply be due to this pattern of isochore degradation. In which case we can no longer reject BGC.

However, this re-evaluation will need to be more than just repeating what has been done before. Models of sequence evolution which allow for variation in substitution patterns and for degradation of isochore patterns need to be developed in order to see how such variation may affect patterns of base composition. This will allow us to develop predictions of what is expected under neutral models. The question of what causes the variation commonly

seen in mutation rates remains and the debate between the various theories used to explain the variation continues.

Bielawski, J.P., Dunn, K.A. and Yang, Z.H.: Rates of nucleotide substitution and mammalian nuclear gene evolution: Approximate and maximum-likelihood methods lead to different conclusions. Genetics 156 (2000) 1299-1308.

Blumenthal, T.: Gene clusters and polycistronic transcription in eukaryotes. BioEssays 20 (1998) 480-487.

Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M. and Kim, S.K.: A global analysis of Caenorhabditis elegans operons. Nature 417 (2002) 851-854.

Castresana, J.: Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. Nucleic Acids Research 30 (2002) 1751-1756.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D. and Galtier, N.: Vanishing GC-rich isochores in mammalian genomes. In: Bernardi, G. (Ed.), Molecular evolution: evolution, genomics and bioinformatics, Sorrento(Naples), 2002, pp. 65.

Eyre-Walker, A.: Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. Genetics 152 (1999) 675-683.

Goldman, N. and Yang, Z.H.: Codon-based model of nucleotide substitution for protein-coding dna sequences. Molecular Biology and Evolution 11 (1994) 725-736.

Hill, W.G. and Robertson, A.: The effect of linkage on limits to artificial selection. Genet Res 8 (1966) 269-94.

Kruglyak, S. and Tang, S.: Regulation of adjacent yeast genes. Trends in Genetics 16 (2000) 109-111.

Kumar, S. and Subramanian, S.: Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America 99 (2002) 803-808.

Lercher, M.J., Urrutia, A.O. and Hurst, L.D.: Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nature Genetics 31 (2002) 180-183.

Pesole, G., Dellisanti, G., Preparata, G. and Saccone, C.: The importance of base composition in the correct assessment of genetic distance. Journal of Molecular Evolution 41 (1995) 1124-1127.

Robinson, M., Gautier, C. and Mouchiroud, D.: Evolution of isochores in rodents. Molecular Biology and Evolution 14 (1997) 823-828.

Smith, N.G.C. and Eyre-Walker, A.: Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Molecular Biology and Evolution 18 (2001) 982-986.

Smith, N.G.C. and Hurst, L.D.: Sensitivity of patterns of molecular evolution to alterations in methodology: a critique of Hughes and Yeager. J. Mol. Evol. 47 (1998) 493-500.

Tamura, K.: Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+c-content biases. Molecular Biology and Evolution 9 (1992) 678-687.

Appendix A: Dataset used in Chapter 1: The proteins of Linked genes evolve at similar rates.

Mouse	Rat	chr.	cM pos.	ks	ka	ka/ks	cpgks	cpgka	GC4	duoks	duoka
L11065	U00442	1	5.5	0.185	0.006	0.03243	0.183	0.006	0.597	0.183	0.005
M20658	M95578	1	19.5	0.273	0.08	0.29304	0.242	0.077	0.458	0.238	0.054
X59769	Z22812	1	19.5	0.162	0.054	0.33333	0.144	0.054	0.522	0.141	0.037
Y07519	U04319	1	20.6	0.177	0.106	0.59887	0.158	0.108	0.407	0.147	0.078
X81323	U50194	1	27.0	0.101	0.004	0.0396	0.087	0.004	0.303	0.101	0.004
M34563	X55288	1	30.1	0.279	0.055	0.19713	0.22	0.053	0.56	0.239	0.035
X05719	U37121	1	30.1	0.137	0.046	0.33577	0.119	0.042	0.491	0.126	0.041
u46027	X14788	1	31.0	0.124	0.027	0.21774	0.12	0.027	0.331	0.104	0.008
L12447	M62781	1	36.1	0.101	0.003	0.0297	0.074	0.003	0.754	0.101	0.003
X12521	M17096	1	38.4	0.079	0.017	0.21519	0.079	0.017	0.609	0.08	0
M28383	J05167	1	41.0	0.168	0.009	0.05357	0.145	0.009	0.667	0.158	0.006
U11812	X92563	1	41.0	0.155	0.039	0.25161	0.142	0.039	0.616	0.128	0.026
X69618	M36453	1	41.6	0.186	0.027	0.14516	0.181	0.027	0.558	0.173	0.018
X70296	A03913	1	48.6	0.195	0.035	0.17949	0.173	0.036	0.534	0.175	0.025
U09930	J02612	1	51.7	0.191	0.032	0.16754	0.161	0.033	0.562	0.179	0.028
L10076	X74835	1	52.3	0.173	0.024	0.13873	0.148	0.022	0.712	0.163	0.016
m24086	m60737	1	53.6	0.172	0.015	0.08721	0.155	0.015	0.667	0.16	0.012
af000236	aj010828	1	55.6	0.257	0.014	0.05447	0.222	0.014	0.774	0.244	0.01
L31532	L14680	1	59.8	0.091	0.024	0.26374	0.091	0.024	0.719	0.083	0.01
X16490	X64563	1	61.1	0.169	0.064	0.3787	0.143	0.065	0.442	0.142	0.043
l41366	af039583	1	67.6	0.234	0.185	0.7906	0.223	0.188	0.34	0.179	0.122
U06431	X15741	1	68.5	0.176	0.075	0.42614	0.158	0.075	0.59	0.133	0.049
x97399	D38104	1	69.1	0.161	0.036	0.2236	0.146	0.035	0.594	0.154	0.026
X15784	M24393	1	72.3	0.092	0.027	0.29348	0.08	0.027	0.661	0.081	0.017
M88242	L25925	1	76.2	0.21	0.02	0.09524	0.182	0.02	0.5	0.2	0.015
M87861	L23088	1	86.6	0.228	0.054	0.23684	0.2	0.054	0.516	0.189	0.039
X16646	J02701	1	86.8	0.171	0.022	0.12865	0.144	0.02	0.67	0.171	0.019
J04967	D13555	1	87.2	0.148	0.034	0.22973	0.118	0.035	0.616	0.121	0.014
M62860	K03242	1	92.4	0.111	0.013	0.11712	0.107	0.013	0.715	0.111	0.013
M79361	X03468	1	92.6	0.35	0.155	0.44286	0.324	0.154	0.465	0.274	0.084
X53526	X13016	1	93.3	0.176	0.209	1.1875	0.162	0.207	0.421	0.086	0.136
X17496	M83176	1	94.2	0.157	0.14	0.89172	0.13	0.144	0.46	0.098	0.078
Y00426	X55761	1	94.2	0.155	0.123	0.79355	0.145	0.124	0.439	0.114	0.101
x14206	u94340	1	98.6	0.268	0.029	0.10821	0.236	0.029	0.64	0.231	0.014
K02891	M55049	2	6.4	0.19	0.1	0.52632	0.181	0.101	0.495	0.14	0.066
L16980	M72422	2	9.0	0.18	0.009	0.05	0.146	0.008	0.534	0.172	0.005
D10028	X63255	2	12.0	0.152	0.001	0.00658	0.121	0.001	0.653	0.15	0.001
S50200	L12407	2	15.5	0.204	0.074	0.36275	0.166	0.076	0.65	0.176	0.048
u33169#	m69157#	2	16.0	0.193	0.043	0.2228	0.161	0.044	0.609	0.155	0.021
X66223	L06482	2	17.0	0.17	0.007	0.04118	0.147	0.007	0.676	0.17	0.007
M31690	X12459	2	20.0	0.162	0.009	0.05556	0.124	0.009	0.677	0.154	0.009
X14607	X13295	2	27.0	0.188	0.105	0.55851	0.166	0.108	0.667	0.165	0.082
M34141	S67721	2	29.0	0.181	0.029	0.16022	0.144	0.029	0.679	0.164	0.02
M65287	S48190	2	30.0	0.105	0.001	0.00952	0.097	0.001	0.34	0.105	0.001
S53744	L08595	2	34.5	0.05	0.007	0.14	0.047	0.007	0.644	0.045	0.003
X58384	J04591	2	35.0	0.168	0.043	0.25595	0.135	0.044	0.398	0.144	0.029
L00993	X67859	2	41.0	0.106	0.02	0.18868	0.09	0.019	0.282	0.097	0.019
m17640	x74832	2	44.0	0.178	0.009	0.05056	0.148	0.01	0.706	0.168	0.003
d14883	af049882	2	49.6	0.206	0.049	0.23786	0.155	0.05	0.712	0.175	0.024
M35523	U23407	2	54.0	0.124	0.004	0.03226	0.102	0.004	0.621	0.124	0.004
U11763	X67857	2	54.0	0.078	0.028	0.35897	0.065	0.029	0.658	0.072	0.028
X52108	M11670	2	57.0	0.183	0.026	0.14208	0.157	0.027	0.491	0.171	0.015
u12932	m27048	2	60.0	0.243	0.014	0.05761	0.199	0.014	0.543	0.243	0.014
U03723	X16002	2	61.0	0.136	0.008	0.05882	0.116	0.008	0.569	0.127	0.001
X55573	M61175	2	62.0	0.08	0.004	0.05	0.069	0.004	0.597	0.08	0.004
X15830	M63901	2	64.0	0.116	0.012	0.10345	0.092	0.013	0.409	0.112	0.011
U18933	D37880	2	67.1	0.126	0.015	0.11905	0.092	0.015	0.623	0.117	0.012
X57437	M29591	2	71.0	0.173	0.05	0.28902	0.15	0.047	0.597	0.155	0.035
M88355	M25649	2	73.0	0.197	0.013	0.06599	0.197	0.013	0.794	0.197	0.013

M36033	L01702	2	74.0	0.148	0.014	0.09459	0.133	0.013	0.544	0.129	0.008
S93521	D16237	2	74.6	0.161	0.042	0.26087	0.143	0.043	0.619	0.154	0.028
m13685	d50093	2	75.2	0.158	0.016	0.10127	0.123	0.016	0.661	0.152	0.011
x51429	af019974	2	75.6	0.167	0.071	0.42515	0.152	0.071	0.552	0.138	0.049
L25602	Z25868	2	76.1	0.18	0.023	0.12778	0.164	0.024	0.656	0.16	0.014
M55669	M76706	2	81.4	0.186	0.003	0.01613	0.154	0.003	0.595	0.186	0.003
u26176	u04738	2	84.0	0.152	0.021	0.13816	0.131	0.02	0.707	0.134	0.013
M64228	X16476	2	97.0	0.195	0.014	0.07179	0.171	0.014	0.803	0.189	0.01
x74983	x70667	2	100.0	0.223	0.036	0.16143	0.201	0.036	0.793	0.188	0.017
u32330	s39779	2	104.0	0.131	0.091	0.69466	0.117	0.088	0.496	0.114	0.076
U09421	L33916	2	110.0	0.163	0.005	0.03067	0.147	0.005	0.571	0.156	0.001
K00811	X58294	3	10.5	0.143	0.042	0.29371	0.121	0.043	0.5	0.13	0.018
j05118	u67914	3	13.2	0.168	0.032	0.19048	0.152	0.032	0.384	0.149	0.028
K02109	u75581	3	13.9	0.226	0.036	0.15929	0.184	0.037	0.672	0.221	0.024
x16986	j03145	3	14.4	0.175	0.026	0.14857	0.151	0.027	0.496	0.148	0.018
M81591	M15944	3	29.6	0.156	0.008	0.05128	0.144	0.008	0.404	0.156	0.008
U07159	J02791	3	40.1	0.329	0.02	0.06079	0.275	0.021	0.5	0.317	0.016
M16465	J03627	3	41.7	0.131	0.025	0.19084	0.119	0.025	0.611	0.131	0.025
X16190	J03628	3	41.7	0.038	0.018	0.47368	0.028	0.018	0.636	0.038	0.018
X53802	M58587	3	42.1	0.132	0.061	0.46212	0.114	0.061	0.641	0.12	0.053
x94444	af010306	3	42.7	0.242	0.029	0.11983	0.202	0.03	0.559	0.239	0.025
K02060	X06483	3	43.4	0.112	0	0	0.083	0	0.727	0.112	0
x61675	m83092	3	45.6	0.205	0.036	0.17561	0.178	0.035	0.682	0.184	0.02
M18934	X05111	3	46.2	0.193	0.119	0.61658	0.183	0.118	0.564	0.133	0.079
L31932	M74535	3	47.6	0.141	0.01	0.07092	0.116	0.01	0.644	0.139	0.008
M63695	D26439	3	48.0	0.21	0.087	0.41429	0.194	0.088	0.567	0.153	0.055
m54943	m13897	3	48.5	0.177	0.054	0.30508	0.129	0.057	0.556	0.166	0.042
x97227	m57276	3	48.5	0.173	0.038	0.21965	0.168	0.038	0.531	0.148	0.025
M30440	X16003	3	48.8	0.079	0	0	0.061	0	0.591	0.079	0
M58567	M38178	3	49.1	0.172	0.068	0.39535	0.158	0.069	0.521	0.142	0.042
M26071	U07619	3	50.0	0.215	0.101	0.46977	0.186	0.102	0.559	0.162	0.061
L32178	M59742	3	50.4	0.112	0.002	0.01786	0.078	0.002	0.72	0.112	0.001
M30441	X16001	3	52.3	0.153	0.004	0.02614	0.138	0.005	0.727	0.151	0.004
M65034	M35992	3	55.0	0.219	0.035	0.15982	0.157	0.037	0.389	0.179	0.029
L28836	D90038	3	56.6	0.195	0.014	0.07179	0.185	0.014	0.435	0.175	0.009
L28177	L32591	3	70.5	0.25	0.018	0.072	0.228	0.018	0.756	0.237	0.007
u20257	x98746	3	71.2	0.27	0.037	0.13704	0.218	0.036	0.533	0.247	0.026
D17433	D28581	3	75.8	0.348	0.017	0.04885	0.309	0.018	0.521	0.332	0.011
af024621	d38494	4	10.5	0.174	0.029	0.16667	0.148	0.03	0.704	0.17	0.024
j00643	v01253	4	10.5	0.017	0.009	0.52941	0.017	0.009	0.52	0.017	0.009
M21531	M27839	4	10.5	0.185	0.004	0.02162	0.163	0.004	0.372	0.185	0.004
U17985	X55812	4	13.9	0.159	0.003	0.01887	0.124	0.003	0.653	0.159	0.003
X77585	X14878	4	24.6	0.188	0.007	0.03723	0.146	0.008	0.614	0.188	0.007
D28812	S87544	4	30.6	0.189	0.066	0.34921	0.161	0.066	0.684	0.18	0.05
X13752	X04959	4	30.6	0.121	0.014	0.1157	0.088	0.014	0.58	0.113	0.006
M27008	J00696	4	31.4	0.246	0.155	0.63008	0.235	0.152	0.568	0.175	0.124
D37801	U17971	4	38.0	0.219	0.018	0.08219	0.193	0.018	0.638	0.203	0.013
k00020	d87919	4	42.6	0.149	0.137	0.91946	0.144	0.138	0.646	0.126	0.087
M23384	M13979	4	52.0	0.151	0.01	0.06623	0.117	0.009	0.688	0.15	0.007
M91458	M62763	4	52.0	0.251	0.048	0.19124	0.215	0.045	0.506	0.208	0.03
X67056	M88595	4	52.0	0.183	0.018	0.09836	0.147	0.018	0.707	0.159	0.005
L05516	M95493	4	57.0	0.139	0.058	0.41727	0.123	0.059	0.563	0.1	0.031
m91443	x76168	4	57.5	0.213	0.021	0.09859	0.172	0.02	0.709	0.203	0.011
X14961	J02773	4	61.0	0.277	0.032	0.11552	0.236	0.033	0.629	0.213	0.019
L11064	D16348	4	64.8	0.164	0.015	0.09146	0.15	0.015	0.828	0.156	0.013
M60523	D10864	4	66.0	0.114	0.004	0.03509	0.106	0.004	0.712	0.114	0.004
u18119	u07798	4	68.0	0.177	0.05	0.28249	0.153	0.051	0.776	0.127	0.023
u28244	m37127	4	68.0	0.199	0.108	0.54271	0.173	0.111	0.605	0.169	0.08
u66873	u03763	4	68.0	0.132	0.046	0.34848	0.132	0.046	0.69	0.126	0.044
J02980	J03572	4	70.2	0.259	0.015	0.05792	0.226	0.015	0.68	0.239	0.008

d16497	m25297	4	76.5	0.15	0.119	0.79333	0.15	0.119	0.636	0.077	0.102
k02781	m27498	4	76.5	0.145	0.038	0.26207	0.105	0.036	0.633	0.121	0.016
X52379	X02610	4	79.0	0.17	0.02	0.11765	0.126	0.02	0.659	0.156	0.013
Z21674	X17037	4	79.4	0.107	0.043	0.40187	0.103	0.043	0.518	0.09	0.037
J03398	L15079	5	1.0	0.202	0.014	0.06931	0.171	0.014	0.531	0.197	0.011
D10213	D90102	5	4.0	0.105	0.006	0.05714	0.094	0.006	0.384	0.102	0.005
D29678	L02121	5	12.0	0.126	0.002	0.01587	0.109	0.002	0.637	0.12	0.002
J03783	M26744	5	17.0	0.138	0.071	0.51449	0.134	0.072	0.412	0.121	0.055
M13019	L12138	5	18.0	0.153	0.014	0.0915	0.143	0.015	0.526	0.146	0.01
X83971	U18913	5	18.0	0.203	0.012	0.05911	0.159	0.013	0.694	0.193	0.006
m99376	x57659	5	20.0	0.095	0.002	0.02105	0.08	0.002	0.807	0.089	0.001
l20330	m69118	5	23.0	0.18	0.021	0.11667	0.167	0.018	0.732	0.161	0.015
L11332	D29646	5	28.0	0.177	0.069	0.38983	0.163	0.07	0.569	0.155	0.045
u38261	z24721	5	31.0	0.242	0.107	0.44215	0.212	0.106	0.733	0.171	0.053
d85605	d50608	5	34.0	0.13	0.02	0.15385	0.108	0.021	0.695	0.115	0.014
Y00864	D12524	5	42.0	0.216	0.039	0.18056	0.181	0.039	0.615	0.194	0.028
L01119	U00935	5	44.9	0.223	0.021	0.09417	0.213	0.022	0.552	0.213	0.019
M10114	K02598	5	45.0	0.153	0.129	0.84314	0.129	0.126	0.357	0.142	0.095
x13484	j00711	5	45.0	0.12	0.101	0.84167	0.106	0.102	0.434	0.109	0.078
D12648	X55183	5	51.0	0.249	0.091	0.36546	0.197	0.088	0.5	0.188	0.054
J04596	d11444	5	51.0	0.157	0.057	0.36306	0.158	0.057	0.704	0.081	0.028
M86829	U17035	5	53.0	0.329	0.129	0.3921	0.32	0.13	0.585	0.276	0.08
u27267	u90448	5	53.0	0.241	0.115	0.47718	0.222	0.112	0.582	0.196	0.097
u64827	u27562	5	55.0	0.248	0.062	0.25	0.219	0.059	0.469	0.21	0.046
u59283	d13121	5	56.0	0.065	0.041	0.63077	0.065	0.041	0.677	0.053	0.033
X16151	M14656	5	56.0	0.203	0.088	0.4335	0.181	0.09	0.453	0.163	0.051
D00611	X54640	5	60.0	0.208	0.082	0.39423	0.17	0.085	0.66	0.177	0.053
M60559	X16072	5	60.0	0.203	0	0	0.159	0	0.709	0.203	0
d14552	u67309	5	65.0	0.201	0.01	0.04975	0.167	0.011	0.722	0.191	0.007
L11163	J05030	5	65.0	0.267	0.015	0.05618	0.214	0.015	0.697	0.268	0.013
M57966	X54423	5	65.0	0.186	0.004	0.02151	0.169	0.004	0.705	0.177	0.002
M28541	M13962	5	72.0	0.235	0.066	0.28085	0.195	0.066	0.653	0.197	0.047
U14166	D31873	5	78.0	0.108	0.014	0.12963	0.086	0.012	0.672	0.108	0.007
d17571	m10068	5	79.0	0.212	0.009	0.04245	0.203	0.009	0.728	0.201	0.005
U03560	M86389	5	79.0	0.236	0.017	0.07203	0.215	0.017	0.709	0.221	0.007
X56518	S50879	5	80.0	0.095	0.01	0.10526	0.083	0.009	0.652	0.093	0.008
M29464	L06238	5	84.0	0.074	0.013	0.17568	0.06	0.013	0.753	0.06	0.005
M26687	D67087	5	88.0	0.168	0.023	0.1369	0.132	0.023	0.676	0.152	0.013
U18542	L13041	6	4.5	0.222	0.037	0.16667	0.199	0.037	0.573	0.213	0.028
U18812	D45862	6	10.5	0.114	0.016	0.14035	0.097	0.016	0.72	0.104	0.012
D90225	M55601	6	13.5	0.165	0	0	0.149	0	0.576	0.137	0
L18868	D28773	6	20.5	0.116	0.068	0.58621	0.103	0.068	0.589	0.094	0.052
L02914	L07268	6	27.0	0.15	0.013	0.08667	0.12	0.014	0.696	0.145	0.005
M17534	X04310	6	30.5	0.193	0.139	0.72021	0.193	0.139	0.634	0.132	0.096
M92420	M31076	6	35.8	0.089	0.008	0.08989	0.082	0.009	0.67	0.089	0.008
L03292	M96601	6	38.5	0.092	0.007	0.07609	0.07	0.007	0.64	0.088	0.007
af053471	j03754	6	49.5	0.139	0.001	0.00719	0.107	0.001	0.717	0.136	0.001
l42198	j03960	6	53.2	0.2	0.014	0.07	0.163	0.013	0.671	0.183	0.007
M32599	M17701	6	56.0	0.167	0.01	0.05988	0.118	0.01	0.698	0.158	0.007
l08115	x76489	6	57.0	0.166	0.032	0.19277	0.142	0.032	0.648	0.163	0.02
M60468	M63122	6	57.1	0.184	0.092	0.5	0.17	0.092	0.645	0.156	0.062
M83749	L09752	6	60.0	0.101	0.006	0.05941	0.073	0.007	0.717	0.098	0.005
M96688	X17621	6	60.0	0.143	0.006	0.04196	0.12	0.006	0.678	0.143	0.006
X53257	M34643	6	60.0	0.112	0.004	0.03571	0.105	0.004	0.607	0.112	0.004
Y00305	X12589	6	60.0	0.071	0.003	0.04225	0.066	0.003	0.716	0.071	0.003
X52380	M11931	6	60.2	0.156	0.008	0.05128	0.113	0.008	0.684	0.147	0.005
M25389	J04544	6	62.0	0.208	0.054	0.25962	0.198	0.055	0.54	0.179	0.037
X51905	U07181	6	62.0	0.269	0.004	0.01487	0.209	0.005	0.688	0.26	0.003
D10651	U11419	6	64.5	0.138	0.002	0.01449	0.11	0.002	0.648	0.137	0.002
D26157	D28966	7	2.5	0.187	0.041	0.21925	0.159	0.041	0.687	0.165	0.024

Z22661	X15512	7	4.0	0.313	0.093	0.29712	0.285	0.095	0.732	0.274	0.04
D10011	U08258	7	6.5	0.122	0.003	0.02459	0.1	0.003	0.764	0.121	0.002
aj009862	x52498	7	6.5	0.118	0.007	0.05932	0.098	0.007	0.56	0.108	0.004
X74938	L09648	7	6.5	0.17	0.016	0.09412	0.155	0.015	0.624	0.16	0.012
U09181	X58499	7	9.0	0.209	0.004	0.01914	0.2	0.004	0.635	0.209	0.004
M58691	X63369	7	10.2	0.142	0.018	0.12676	0.126	0.018	0.578	0.134	0.009
M31811	M16800	7	11.0	0.165	0.013	0.07879	0.141	0.013	0.726	0.15	0.006
af024570	aj222691	7	23.0	0.201	0.02	0.0995	0.161	0.02	0.705	0.184	0.01
L17022	X14323	7	23.0	0.235	0.047	0.2	0.215	0.048	0.59	0.198	0.029
L22472	U49729	7	23.0	0.031	0.01	0.32258	0.026	0.01	0.717	0.021	0.005
u25145	j00749	7	23.0	0.141	0.014	0.09929	0.126	0.014	0.67	0.117	0.008
J04758	X53501	7	23.5	0.197	0.02	0.10152	0.171	0.021	0.638	0.192	0.018
m17587	u07177	7	23.5	0.22	0.064	0.29091	0.178	0.063	0.491	0.182	0.038
U13687	X01964	7	23.5	0.254	0.014	0.05512	0.224	0.015	0.615	0.226	0.008
X51528	X68282	7	25.0	0.182	0.009	0.04945	0.173	0.009	0.705	0.179	0.007
X62648	J05497	7	27.6	0.103	0.011	0.1068	0.099	0.009	0.291	0.095	0.006
x59300	m81142	7	28.2	0.116	0.005	0.0431	0.112	0.005	0.537	0.114	0.004
l37663	s53987	7	30.0	0.111	0.005	0.04505	0.092	0.005	0.65	0.097	0.002
X75313	U03390	7	31.2	0.18	0.001	0.00556	0.16	0.001	0.59	0.171	0.001
L26489	X55660	7	39.0	0.184	0.016	0.08696	0.154	0.015	0.676	0.17	0.01
M84145	M77694	7	42.5	0.172	0.012	0.06977	0.155	0.012	0.546	0.17	0.01
u94593	af039033	7	50.0	0.162	0.005	0.03086	0.139	0.005	0.641	0.159	0.003
M27959	X69903	7	52.0	0.178	0.126	0.70787	0.166	0.126	0.588	0.119	0.069
L06144	U10188	7	59.0	0.173	0.017	0.09827	0.152	0.018	0.716	0.167	0.014
U08439	X12554	7	61.0	0.221	0.014	0.06335	0.167	0.015	0.642	0.221	0.014
M84524	X54862	7	66.0	0.131	0.039	0.29771	0.1	0.037	0.557	0.104	0.026
M14951	X14834	7	69.0	0.032	0.016	0.5	0.027	0.014	0.684	0.027	0.014
x04724	j00748	7	69.0	0.152	0.027	0.17763	0.123	0.028	0.695	0.144	0.022
M69200	L22651	7	69.2	0.156	0.015	0.09615	0.116	0.015	0.642	0.15	0.013
S45012	U19894	7	69.2	0.2	0.11	0.55	0.174	0.107	0.748	0.136	0.038
J04992	M73701	7	70.0	0.116	0.005	0.0431	0.096	0.005	0.696	0.101	0.002
M64403	D14014	7	72.3	0.123	0.012	0.09756	0.106	0.012	0.801	0.112	0.007
M34163	X73579	8	0.4	0.252	0.06	0.2381	0.223	0.062	0.59	0.232	0.047
J05149	M29014	8	1.0	0.189	0.005	0.02646	0.157	0.005	0.593	0.181	0.004
M25244	M34959	8	1.0	0.202	0.1	0.49505	0.171	0.103	0.601	0.183	0.064
M64688	M35535	8	4.0	0.182	0.024	0.13187	0.182	0.024	0.715	0.179	0.02
J03520	M23697	8	9.0	0.273	0.041	0.15018	0.229	0.042	0.589	0.249	0.027
d49744	m81225	8	9.5	0.199	0.008	0.0402	0.17	0.006	0.562	0.193	0.005
M58588	M58590	8	26.0	0.191	0.052	0.27225	0.168	0.053	0.498	0.176	0.039
Z46757	D84418	8	31.0	0.254	0.013	0.05118	0.248	0.013	0.521	0.254	0.013
X61232	X51406	8	32.6	0.155	0.014	0.09032	0.136	0.014	0.599	0.129	0.004
L32955	D28508	8	33.0	0.228	0.04	0.17544	0.2	0.04	0.682	0.2	0.024
X14926	X53363	8	37.0	0.106	0.007	0.06604	0.08	0.007	0.61	0.106	0.005
S80191	X65296	8	43.2	0.559	0.158	0.28265	0.54	0.16	0.486	0.361	0.071
M36778	M17526	8	45.5	0.347	0.047	0.13545	0.325	0.048	0.687	0.268	0.014
J05154	X54096	8	53.0	0.222	0.04	0.18018	0.198	0.038	0.656	0.201	0.029
U12961	J02640	8	53.3	0.251	0.031	0.12351	0.214	0.028	0.521	0.232	0.024
M37829	X14209	8	64.0	0.207	0.032	0.15459	0.162	0.033	0.667	0.188	0.022
X51971	U12268	8	66.0	0.202	0.08	0.39604	0.183	0.08	0.708	0.167	0.043
D13139	L07315	8	67.0	0.243	0.067	0.27572	0.209	0.069	0.628	0.201	0.042
D16142	D30035	8	67.0	0.112	0.015	0.13393	0.101	0.015	0.383	0.1	0.009
L28095	U14647	9	1.0	0.173	0.05	0.28902	0.158	0.051	0.346	0.148	0.04
L36244	L24374	9	1.0	0.206	0.071	0.34466	0.181	0.072	0.478	0.174	0.048
S59388	D13566	9	5.0	0.19	0.031	0.16316	0.17	0.03	0.639	0.173	0.022
x64414	x13722	9	5.0	0.234	0.062	0.26496	0.195	0.062	0.684	0.2	0.037
l20334	u10699	9	6.0	0.217	0.026	0.11982	0.153	0.028	0.741	0.173	0.018
M99054	M76110	9	6.0	0.247	0.033	0.1336	0.196	0.031	0.646	0.216	0.017
m55181	m28263	9	7.0	0.272	0.022	0.08088	0.236	0.023	0.689	0.261	0.015
X53953	L20681	9	15.0	0.117	0.01	0.08547	0.093	0.01	0.598	0.109	0.01
x71788	x71463	9	25.0	0.3	0.028	0.09333	0.279	0.027	0.686	0.272	0.019

Y00635	S79711	9	26.0	0.14	0.108	0.77143	0.128	0.11	0.444	0.105	0.07
M64249	M00002	9	27.0	0.194	0.09	0.46392	0.184	0.091	0.8	0.157	0.063
X64262	X00558	9	27.0	0.295	0.185	0.62712	0.255	0.186	0.752	0.184	0.108
X55674	X56065	9	28.0	0.113	0	0	0.084	0	0.703	0.113	0
M63170	S77138	9	29.0	0.164	0.008	0.04878	0.141	0.005	0.636	0.152	0.005
D00659	M33986	9	31.0	0.209	0.045	0.21531	0.183	0.046	0.599	0.183	0.027
U05247	X58631	9	32.0	0.107	0.005	0.04673	0.089	0.005	0.662	0.105	0.002
L02526	Z16415	9	36.0	0.137	0.001	0.0073	0.115	0.001	0.621	0.133	0.001
M14044	X66871	9	37.0	0.215	0.012	0.05581	0.178	0.012	0.667	0.202	0.011
X58426	M16235	9	39.0	0.195	0.06	0.30769	0.166	0.059	0.647	0.152	0.038
X64840	L09656	9	42.0	0.082	0.004	0.04878	0.073	0.004	0.359	0.08	0.004
M85151	X62944	9	46.0	0.067	0.01	0.14925	0.064	0.01	0.77	0.067	0.01
M73483	X78848	9	48.0	0.278	0.076	0.27338	0.25	0.078	0.629	0.257	0.045
L20899	X67241	9	50.0	0.231	0.05	0.21645	0.208	0.051	0.678	0.204	0.034
U06119	Y00708	9	50.0	0.178	0.037	0.20787	0.149	0.038	0.589	0.154	0.026
u59761	d84450	9	51.0	0.161	0.086	0.53416	0.145	0.086	0.359	0.134	0.056
S69114	S67770	9	52.0	0.22	0.032	0.14545	0.189	0.031	0.778	0.174	0.013
X60367	M16459	9	52.0	0.237	0.004	0.01688	0.189	0.004	0.681	0.237	0.004
j03299	d38380	9	56.0	0.234	0.083	0.3547	0.201	0.086	0.635	0.159	0.047
X74154	M13949	9	57.0	0.212	0.011	0.05189	0.174	0.008	0.778	0.201	0
X78936	M77184	9	58.0	0.129	0.007	0.05426	0.105	0.007	0.666	0.121	0.005
M13963	M17528	9	59.0	0.178	0.005	0.02809	0.158	0.005	0.723	0.174	0.001
u28406	af003954	9	72.0	0.143	0.042	0.29371	0.131	0.042	0.484	0.133	0.031
M60320	M26686	10	7.0	0.197	0.008	0.04061	0.155	0.007	0.427	0.182	0.004
AF06275	D16349	10	8.0	0.222	0.024	0.10811	0.179	0.025	0.61	0.202	0.014
M58661	Z11663	10	26.0	0.187	0.053	0.28342	0.162	0.054	0.549	0.136	0.035
X61576	X06656	10	29.0	0.19	0.001	0.00526	0.147	0.001	0.69	0.19	0.001
J05277	J04526	10	30.0	0.197	0.018	0.09137	0.158	0.018	0.721	0.185	0.013
X06746	D83508	10	35.0	0.187	0.006	0.03209	0.17	0.006	0.512	0.175	0.005
L10613	S73424	10	40.9	0.118	0.004	0.0339	0.091	0.005	0.814	0.118	0.004
J03928	X58865	10	41.5	0.212	0.004	0.01887	0.161	0.004	0.685	0.204	0.004
D10849	D32080	10	43.0	0.296	0.042	0.14189	0.252	0.044	0.78	0.246	0.024
M11768	S73894	10	43.0	0.261	0.099	0.37931	0.204	0.105	0.644	0.198	0.048
X04573	V01233	10	43.0	0.198	0.092	0.46465	0.141	0.096	0.697	0.147	0.044
X52101	X74565	10	43.0	0.192	0.008	0.04167	0.165	0.008	0.623	0.186	0.005
X73361	L14462	10	43.0	0.135	0	0	0.111	0	0.739	0.135	0
X51942	M12337	10	47.0	0.137	0.011	0.08029	0.12	0.011	0.525	0.122	0.008
Z30970	U27201	10	47.0	0.1	0.004	0.04	0.084	0.004	0.72	0.089	0.002
x04480	x06043	10	48.0	0.16	0.009	0.05625	0.141	0.009	0.556	0.154	0.005
L07645	M58308	10	51.0	0.183	0.006	0.03279	0.143	0.007	0.55	0.17	0.004
k00083	af010466	10	67.0	0.291	0.075	0.25773	0.247	0.075	0.646	0.221	0.063
u96386	af140032	10	69.0	0.132	0.036	0.27273	0.119	0.036	0.619	0.118	0.02
M37897	X60675	10	69.9	0.173	0.077	0.44509	0.173	0.077	0.661	0.13	0.057
X12810	M32748	11	0.25	0.247	0.052	0.21053	0.213	0.051	0.689	0.219	0.033
X81579	M58634	11	1.3	0.181	0.024	0.1326	0.143	0.021	0.558	0.177	0.023
X81581	M31837	11	1.35	0.192	0.037	0.19271	0.165	0.038	0.792	0.146	0.021
Y00094	J02998	11	11.0	0.201	0.004	0.0199	0.191	0.004	0.379	0.201	0.004
U10420	X56420	11	16.0	0.163	0.042	0.25767	0.142	0.043	0.577	0.141	0.027
V00714	M17083	11	16.0	0.218	0.088	0.40367	0.167	0.093	0.675	0.115	0.038
m86566	I08490	11	19.0	0.149	0	0	0.132	0	0.446	0.149	0
m86572	I08497	11	19.0	0.202	0.006	0.0297	0.181	0.006	0.485	0.2	0.005
y12738	I08610	11	19.0	0.163	0.013	0.07975	0.142	0.013	0.782	0.158	0.01
L07037	D16302	11	25.0	0.144	0.013	0.09028	0.124	0.014	0.583	0.136	0.01
X13460	X86086	11	29.5	0.148	0.009	0.06081	0.12	0.009	0.648	0.145	0.008
X04017	D28875	11	29.9	0.126	0.014	0.11111	0.106	0.014	0.704	0.114	0.009
L18888	L18889	11	30.0	0.125	0.009	0.072	0.103	0.009	0.336	0.119	0.005
S73717	D00833	11	30.0	0.159	0.002	0.01258	0.124	0.002	0.649	0.154	0.001
X57497	M38060	11	31.0	0.142	0.004	0.02817	0.115	0.003	0.634	0.139	0.002
u12785	j03637	11	33.0	0.222	0.049	0.22072	0.181	0.048	0.648	0.192	0.032
X80417	X78461	11	33.0	0.173	0.007	0.04046	0.121	0.007	0.719	0.168	0.006

M32240	X62431	11	34.5	0.164	0.013	0.07927	0.14	0.01	0.726	0.158	0.01
L41933	L36029	11	37.0	0.178	0.023	0.12921	0.158	0.023	0.621	0.163	0.019
X53042	J02762	11	37.0	0.206	0.084	0.40777	0.174	0.087	0.658	0.157	0.054
X01237	X13058	11	39.0	0.225	0.056	0.24889	0.195	0.055	0.555	0.2	0.041
M14537	X74833	11	40.0	0.162	0.018	0.11111	0.137	0.018	0.644	0.146	0.013
M23383	D28561	11	40.0	0.137	0.013	0.09489	0.121	0.013	0.63	0.128	0.009
U04331	S69383	11	40.0	0.175	0.032	0.18286	0.154	0.031	0.623	0.156	0.02
X16645	J04629	11	40.0	0.197	0.013	0.06599	0.162	0.014	0.648	0.194	0.012
X57747	Y00979	11	42.0	0.21	0.019	0.09048	0.163	0.019	0.591	0.201	0.007
S72681	M25758	11	44.1	0.104	0.002	0.01923	0.09	0.002	0.535	0.104	0.002
S72408	D44481	11	44.2	0.037	0.003	0.08108	0.033	0.003	0.623	0.03	0.001
u26426	u96637	11	47.0	0.062	0.025	0.40323	0.062	0.025	0.667	0.062	0.017
X70675	U06436	11	47.0	0.115	0.023	0.2	0.115	0.023	0.625	0.117	0.011
X12531	U22414	11	47.59	0.144	0.053	0.36806	0.131	0.053	0.578	0.099	0.042
S68107	S71523	11	48.0	0.163	0.032	0.19632	0.14	0.033	0.619	0.135	0.001
u34295	I08831	11	56.0	0.153	0.032	0.20915	0.134	0.033	0.531	0.153	0.032
X51983	M18028	11	57.0	0.083	0.003	0.03614	0.056	0.004	0.734	0.083	0.003
M86736	M97750	11	60.0	0.168	0.071	0.42262	0.143	0.072	0.575	0.145	0.054
U08378	X91810	11	60.5	0.169	0.001	0.00592	0.14	0.001	0.652	0.165	0.001
X02801	U03700	11	62.0	0.177	0.035	0.19774	0.161	0.035	0.688	0.147	0.014
X72305	L25438	11	62.0	0.163	0.008	0.04908	0.144	0.008	0.779	0.154	0.004
S70439	X15551	11	63.0	0.225	0.096	0.42667	0.186	0.096	0.532	0.18	0.058
M18775	X79321	11	64.0	0.212	0.01	0.04717	0.173	0.01	0.582	0.206	0.008
X02891	J00789	11	65.0	0.123	0.02	0.1626	0.095	0.021	0.775	0.119	0.018
X73052	L48490	11	68.0	0.237	0.003	0.01266	0.179	0.004	0.52	0.237	0.003
X62622	S72594	11	72.0	0.129	0.004	0.03101	0.119	0.004	0.742	0.119	0.002
U07617	X62853	11	75.0	0.11	0.004	0.03636	0.096	0.004	0.728	0.11	0.004
X06453	X02918	11	80.0	0.204	0.018	0.08824	0.166	0.018	0.551	0.19	0.014
X15487	S61865	12	1.0	0.206	0.051	0.24757	0.158	0.051	0.561	0.164	0.029
X60703	X17396	12	15.0	0.166	0.03	0.18072	0.143	0.03	0.565	0.146	0.02
M94623	U04860	12	18.0	0.261	0.043	0.16475	0.238	0.044	0.524	0.227	0.027
Z16406	Z17223	12	20.0	0.134	0.004	0.02985	0.121	0.004	0.597	0.134	0.004
m81831	x62314	12	23.0	0.109	0.007	0.06422	0.085	0.008	0.726	0.104	0.007
u02602	m34842	12	37.0	0.218	0.031	0.1422	0.194	0.03	0.626	0.205	0.022
x80171	I40030	12	39.0	0.172	0.046	0.26744	0.142	0.047	0.647	0.142	0.032
M32745	U03491	12	41.0	0.136	0.004	0.02941	0.126	0.004	0.685	0.136	0.004
M64278	X06832	12	48.0	0.171	0.05	0.2924	0.159	0.049	0.642	0.139	0.038
x69676	m59967	12	53.0	0.206	0.045	0.21845	0.171	0.045	0.858	0.177	0.031
X65687	D30040	12	58.0	0.166	0.003	0.01807	0.131	0.003	0.748	0.161	0.003
X69619	M37482	13	10.0	0.111	0.002	0.01802	0.089	0.002	0.684	0.111	0.002
af011385	I06441	13	14.0	0.24	0.201	0.8375	0.236	0.202	0.293	0.15	0.116
m35662	d21103	13	14.0	0.207	0.114	0.55072	0.18	0.114	0.394	0.159	0.077
x02892	af022935	13	14.0	0.182	0.077	0.42308	0.165	0.078	0.515	0.157	0.056
S37484	X62295	13	16.0	0.214	0.006	0.02804	0.188	0.006	0.69	0.214	0.006
X06086	Y00697	13	30.0	0.184	0.03	0.16304	0.145	0.03	0.561	0.16	0.023
L33878	L13257	13	31.0	0.184	0.01	0.05435	0.162	0.01	0.707	0.163	0.006
M58507	M76705	13	44.0	0.148	0.025	0.16892	0.131	0.025	0.385	0.138	0.018
L03529	M81642	13	47.0	0.208	0.044	0.21154	0.188	0.044	0.783	0.184	0.028
X02389	X65651	14	2.5	0.178	0.062	0.34831	0.145	0.062	0.525	0.153	0.045
D29016	M95591	14	3.0	0.18	0.038	0.21111	0.165	0.039	0.63	0.164	0.026
u40189	z68180	14	10.5	0.159	0.04	0.25157	0.145	0.04	0.667	0.151	0.036
U04672	S75359	14	13.0	0.156	0.006	0.03846	0.141	0.006	0.381	0.156	0.006
I40156	m81231	14	14.0	0.19	0.044	0.23158	0.171	0.045	0.347	0.163	0.029
S48768	M33201	14	14.0	0.154	0.049	0.31818	0.116	0.051	0.465	0.124	0.032
X56848	Z22607	14	14.0	0.136	0.006	0.04412	0.111	0.006	0.642	0.133	0.004
X57024	X14044	14	15.5	0.195	0.005	0.02564	0.162	0.005	0.591	0.191	0.004
D90374	D44495	14	18.5	0.178	0.007	0.03933	0.162	0.007	0.516	0.178	0.007
X04072	M34097	14	20.5	0.197	0.115	0.58376	0.189	0.116	0.543	0.186	0.075
L05670	M16975	14	28.0	0.19	0.038	0.2	0.171	0.039	0.704	0.165	0.026
S69034	X82396	14	28.0	0.187	0.039	0.20856	0.149	0.039	0.491	0.183	0.034

Z37107	X65083	14	32.5	0.147	0.042	0.28571	0.138	0.042	0.5	0.136	0.033
S49542	M30705	14	41.5	0.099	0.012	0.12121	0.08	0.013	0.656	0.095	0.01
u32329	s65355	14	51.0	0.161	0.013	0.08075	0.151	0.013	0.573	0.155	0.009
L13593	M57668	15	4.6	0.156	0.049	0.3141	0.138	0.049	0.408	0.134	0.031
D13458	D28860	15	6.5	0.168	0.01	0.05952	0.143	0.01	0.741	0.162	0.009
M94384	D17469	15	24.7	0.186	0.019	0.10215	0.159	0.018	0.579	0.176	0.015
I17306	j05122	15	43.3	0.159	0.022	0.13836	0.131	0.023	0.664	0.143	0.017
M28052	M55050	15	43.3	0.164	0.1	0.60976	0.155	0.101	0.653	0.133	0.071
X66532	M19036	15	44.9	0.15	0.019	0.12667	0.11	0.019	0.746	0.143	0.016
X59382	M12725	15	45.7	0.091	0.029	0.31868	0.065	0.03	0.66	0.048	0.012
m91000	x63574	15	46.3	0.224	0.019	0.08482	0.202	0.02	0.707	0.209	0.012
S71382	X87635	15	46.7	0.227	0.018	0.0793	0.184	0.018	0.775	0.22	0.01
X57638	M88592	15	48.8	0.196	0.012	0.06122	0.174	0.012	0.649	0.192	0.008
X14943	D38492	15	55.1	0.321	0.006	0.01869	0.267	0.006	0.543	0.31	0.003
D10217	M91561	16	3.4	0.144	0.007	0.04861	0.117	0.007	0.609	0.142	0.005
X61800	M65149	16	9.0	0.206	0.023	0.11165	0.189	0.023	0.826	0.191	0.015
D10939	M64300	16	9.9	0.075	0	0	0.064	0	0.493	0.07	0
U07425	X74550	16	9.9	0.134	0.036	0.26866	0.128	0.037	0.47	0.129	0.022
L34169	D32207	16	13.3	0.107	0.067	0.62617	0.101	0.068	0.48	0.078	0.044
D16106	M18769	16	15.5	0.215	0.037	0.17209	0.198	0.037	0.606	0.187	0.024
x51468	v01271	16	19.0	0.093	0	0	0.062	0	0.705	0.093	0
X82648	X55572	16	21.2	0.173	0.058	0.33526	0.136	0.06	0.58	0.148	0.05
Z12302	U03490	16	22.9	0.118	0.006	0.05085	0.095	0.006	0.519	0.118	0.004
X67274	X53944	16	23.3	0.162	0.014	0.08642	0.139	0.015	0.646	0.156	0.011
M91380	U06864	16	27.3	0.156	0.02	0.12821	0.121	0.02	0.639	0.128	0.013
X60958	U05593	16	28.0	0.261	0.233	0.89272	0.235	0.228	0.434	0.119	0.11
J02809	X06338	16	29.5	0.066	0.004	0.06061	0.053	0.004	0.44	0.066	0.004
D12885	L01506	16	43.5	0.187	0.007	0.03743	0.172	0.007	0.535	0.174	0.005
X06683	Y00404	16	61.0	0.155	0.018	0.11613	0.134	0.018	0.513	0.142	0.011
X60457	M22412	16	64.4	0.206	0.04	0.19417	0.18	0.041	0.688	0.203	0.023
u04710	u59809	17	7.35	0.187	0.034	0.18182	0.157	0.033	0.558	0.173	0.025
X04972	Y00497	17	7.6	0.149	0.031	0.20805	0.122	0.032	0.529	0.136	0.018
L00653	D38455	17	10.4	0.637	0.198	0.31083	0.593	0.198	0.599	0.373	0.076
L28116	U40064	17	13.5	0.183	0.01	0.05464	0.136	0.009	0.824	0.177	0.007
U09507	U24174	17	15.23	0.124	0.031	0.25	0.1	0.032	0.682	0.124	0.031
D38410	M80826	17	17.0	0.265	0.1	0.37736	0.249	0.096	0.595	0.146	0.031
X62742	Z49761	17	18.56	0.229	0.069	0.30131	0.206	0.071	0.65	0.204	0.05
S59862	D10757	17	18.59	0.166	0.025	0.1506	0.128	0.024	0.581	0.161	0.023
M90459	X75306	17	18.6	0.191	0.043	0.22513	0.165	0.044	0.698	0.165	0.029
X14770	X52376	17	18.8	0.173	0.013	0.07514	0.134	0.014	0.839	0.173	0.013
I27086	x77209	17	19.0	0.232	0.006	0.02586	0.203	0.006	0.683	0.226	0.003
X56502	L15619	17	19.02	0.195	0	0	0.161	0	0.536	0.195	0
X02611	X66539	17	19.06	0.157	0.035	0.22293	0.125	0.036	0.7	0.14	0.025
u64572	m99485	17	20.34	0.21	0.032	0.15238	0.186	0.031	0.5	0.2	0.028
S37052	M32167	17	24.2	0.174	0.009	0.05172	0.163	0.009	0.582	0.17	0.007
M74897	S43408	17	30.4	0.18	0.071	0.39444	0.157	0.071	0.58	0.154	0.045
X64361	U39476	17	32.7	0.185	0.009	0.04865	0.157	0.008	0.616	0.183	0.007
M93567	L29281	17	40.0	0.242	0.144	0.59504	0.223	0.147	0.372	0.18	0.073
X62932	J05579	17	45.3	0.217	0.029	0.13364	0.182	0.029	0.593	0.205	0.02
X81143	X93591	17	45.9	0.266	0.031	0.11654	0.225	0.031	0.511	0.244	0.022
M81310	M26199	17	46.5	0.187	0.03	0.16043	0.161	0.031	0.611	0.16	0.022
X61940	X84004	17	50.8	0.173	0.009	0.05202	0.154	0.008	0.744	0.167	0.008
u88623	af144082	18	6.0	0.264	0.014	0.05303	0.211	0.013	0.485	0.253	0.013
D00073	K03252	18	7.0	0.222	0	0	0.186	0	0.58	0.222	0
L15193	M88601	18	8.0	0.218	0.058	0.26606	0.18	0.059	0.444	0.185	0.039
L07264	L05489	18	15.0	0.128	0.041	0.32031	0.11	0.042	0.56	0.128	0.041
M30641	X14232	18	19.0	0.168	0	0	0.123	0	0.716	0.168	0
X04435	M14053	18	20.0	0.171	0.022	0.12865	0.155	0.021	0.432	0.156	0.015
M65142	U11038	18	29.0	0.204	0.022	0.10784	0.175	0.021	0.646	0.192	0.019
I22527	I27081	18	42.0	0.266	0.031	0.11654	0.233	0.032	0.771	0.243	0.019

L09192	U36585	19	0.0	0.148	0.059	0.39865	0.137	0.06	0.656	0.135	0.032
X14309	X89225	19	0.0	0.146	0.077	0.5274	0.135	0.078	0.654	0.113	0.052
M15177	D10728	19	5.0	0.183	0.006	0.03279	0.15	0.006	0.588	0.176	0.003
J05019	M22923	19	8.0	0.176	0.103	0.58523	0.153	0.106	0.425	0.145	0.066
U19265	S79797	19	17.0	0.206	0.033	0.16019	0.185	0.032	0.552	0.184	0.028
X07486	Y00446	19	18.0	0.152	0.03	0.19737	0.134	0.031	0.429	0.138	0.019
U06670	L35767	19	20.0	0.159	0.005	0.03145	0.133	0.005	0.496	0.158	0.005
z27088	a16585	19	21.0	0.208	0.117	0.5625	0.182	0.121	0.47	0.131	0.085
M83649	D26112	19	23.0	0.257	0.24	0.93385	0.237	0.241	0.414	0.174	0.147
D11440	M14103	19	25.0	0.241	0.128	0.53112	0.217	0.128	0.528	0.221	0.076
M30687	L09216	19	29.0	0.175	0.047	0.26857	0.134	0.048	0.542	0.164	0.034
Z23107	X69663	19	33.0	0.12	0.019	0.15833	0.102	0.02	0.806	0.103	0.013
J02623	J04171	19	37.0	0.208	0.028	0.13462	0.175	0.028	0.604	0.181	0.017
M64863	X14086	19	46.0	0.201	0.097	0.48259	0.195	0.098	0.598	0.173	0.06